Advances in the Normal-Normal Hierarchical Model

Advances in the Normal-Normal Hierarchical Model

A dissertation presented

by

Joseph Kelly

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

May 2014

Advances in the Normal-Normal Hierarchical Model

Abstract

This thesis consists of results relating to the theoretical and computational advances in modeling the Normal-Normal hierarchical model.

The first chapter considers the Normal-Normal hierarchical model and proposes a method of estimation known as Gaussian Regression Interactive Multilevel Modeling (GRIMM). The method utilizes Bayesian machinery and chooses priors that lead to good frequency properties and posterior propriety. The method of estimation utilizes a procedure known as Adjustment for Density Maximization (ADM) which allows for estimating the model via differentiation and aims to rectify issues based on maximum likelihood (MLE) and restricted maximum likelihood (REML) estimation that are shown to distort shrinkages and random effect inferences. The advantages and disadvantages for the use of GRIMM are made by comparing inferences based on GRIMM and those on MLE/REML and MCMC. Unlike previous methods GRIMM accounts for the skewness present in the distribution of the random effects and it is shown through simulation and by comparisons with other software such as SAS that this can greatly improve the inference. Whilst not the focus of the chapter the paper notes the existence and makes use of the free R package Rgbp Kelly et al. (2014a) which will allow users to use GRIMM and reproduce the results in the chapter.

The second chapter also considers the estimation of the Normal-Normal hierarchi-

cal model and in particular focuses on the issue of estimation when unequal variances are present in the first level of the model. This chapter proposes a data augmentation scheme that offers some theoretical insights into the problem and for certain data can improve the computational efficiency for estimating the model via Expectation-Maximization based algorithms or by a Monte-Carlo Markov Chain (MCMC). The data augmentation scheme creates missing data such that under the complete data there are equal variances. The new found symmetry under this data augmentation scheme allows for, in some cases, improvements in computational efficiency. A new algorithm combining both the benefits of the data augmentation scheme and adjustment for density maximization is proposed and named Adjustment for Density Maximization Data Augmentation (ADMDA).

Contents

Acknowledgments

Thank you to my advisor, Carl Morris for your constant support, advice and mentorship. Thank you for your contributions to my dissertation and for believing in me and pushing me to become a better statistician. I treasure our friendship and look forward to future collaborations.

I thank the three members of my committee Carl Morris, Joe Blitzstein and Luke Miratrix for their comments and suggestions with regards to the dissertation and my defense. I would also like to thank Don Rubin for his helpful comments and guidance throughout the PhD. A big thank you to my co-author of the Rgbp package, Hyungsuk Tak, not only for his work on the package but for his comments on my dissertation. I would also like to thank my classmates and in particular, Valeria Espinosa and Jon Hennessy, for working with me on various projects and for their friendship and support.

Thank you to Bianca Levy for her constant encouragement, support and laughs throughout the years and thank you to all of my friends and loved ones. Lastly, thank you to my family Vince, Debbie and Jayne Kelly for their never ending love and support.

To Vince, Debbie and Jayne.

Chapter 1

Gaussian Regression Interactive Multilevel Modeling (GRIMM)

Preface

This entirety of this chapter is under the supervision and in collaboration with my advisor Professor Carl Morris.

1.1 Introduction

Gaussian Regression Interactive Multilevel Modeling is a framework and estimation procedure for the Normal-Normal hierarchical model. The Normal-Normal hierarchical model (otherwise known as a multilevel model) is defined at two levels. At

level one it is defined as

$$y_i|\theta_i \sim N(\theta_i, V_i) \text{ for } i = 1 \dots k \tag{1.1}$$

where the $\{V_i\}$ are assumed to be known and often of the form $V_i = \sigma^2/n_i$. Level two is defined as

$$\theta_i|A, \beta \sim N(x_i^{'}\beta, A) \text{ for } i = 1 \dots k. \tag{1.2}$$

where β is a r-dimensional unknown vector and x_i is an r-dimensional known vector of covariate values typically including 1 if an intercept is desired. When there are no covariates present in the model we, thus, define β to be a scalar and $x_i = 1$ for all i.

In Section 1.2 we present the inferential and descriptive versions of this model and then continue in Section 1.4 to assume priors which we show in Section 1.8 to lead to good frequency properties. Posterior propriety for our assumed priors is proved and new posterior propriety results are presented for a class of priors seen in Morris and Lysy (2012).

In Section 1.6 we present the procedure known as Adjustment for Density Maximization (Morris and Tang (2011)) which allows for estimating the model via differentiation and aims to rectify issues based on maximum likelihood (MLE) and restricted maximum likelihood (REML) estimation which we demonstrate distort shrinkages and random effect inferences. We extend the work of Morris and Tang (2011) to incorporate skewness in the approximation of the random effects which we show via simulation and examples to be a desirable extension.

We also compare GRIMM to alternative methods and in Section 1.7 it is shown that in the examples presented GRIMM is comparable to inferences via MCMC. Additionally, in Section 1.8 we evaluate GRIMM's frequency properties and compare with MLE and REML based inferences.

Whilst not the focus of the chapter we note the existence and make use of the free R package, Rgbp, (Kelly et al. (2014a)) which will allow users to use GRIMM and reproduce the results in the chapter and although not presented in the main text a historical perspective of the Normal-Normal hierarchical model is given in Appendix A.

1.2 Descriptive and Inferential Models

In this section we introduce the idea of the descriptive and inferential versions for the Normal-Normal hierarchical model. Table 1.1 builds upon the descriptive models in Section 1.1 and introduces the idea of an equivalent inferential model at Level I and II where in this table $B_i \equiv V_i/(V_i + A)$ is known as the i^{th} shrinkage factor.

Table 1.1: The descriptive and inferential Normal-Normal hierarchical models

	Descriptive	**Inferential**
Level I	$y_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, V_i)$	$y_i \mid \beta, A \sim N(x_i'\beta, V_i + A)$
Level II	$\theta_i \mid \beta, A \overset{ind}{\sim} N(x_i'\beta, A)$	$\theta_i \mid y_i, \beta, A \overset{ind}{\sim} N((1 - B_i)y_i + B_i x_i'\beta, V_i(1 - B_i))$

The inferential Level I model, also known as the marginal distribution of the data conditional on the hyper-parameters, can easily be derived by noting that the marginals of a multivariate Normal are Normal with mean and variance,

$$\begin{aligned} E(y_i \mid \beta, A) &= E(E(y_i \mid \beta, A, \theta_i) \mid \beta, A) \\ &= x_i'\beta \qquad (1.3) \\ \text{Var}(y_i \mid \beta, A) &= E(\text{Var}(y_i \mid \beta, A, \theta_i) \mid \beta, A) + \text{Var}(E(y_i \mid \beta, A, \theta_i) \mid \beta, A) \\ &= V_i + A. \qquad (1.4) \end{aligned}$$

The inferential level II model can be derived by one application of Bayes' rule

$$\begin{aligned} p(\theta_i \mid y_i, A, \beta) &\propto p(y_i \mid \theta_i) p(\theta_i \mid \beta, A) \\ &\propto N(y_i \mid \theta_i, V_i) N(\theta \mid x_i^{'}\beta, A) \\ &\propto N(\theta_i \mid (1 - B_i) y_i + B_i x_i^{'}\beta, V_i(1 - B_i)). \end{aligned} \tag{1.5}$$

where used in this context $N(x \mid \mu, \sigma^2)$ represents the probability density function (pdf) of a Normal with mean μ and variance σ^2 evaluated at x.

It is noted that the descriptive and inferential models are equivalent in that they produce the same joint distribution of the data, $\mathbf{y}$, and random effects, $\Theta \equiv (\theta_1, \ldots, \theta_k)$, conditional on the hyperparameters, (β, A). The descriptive version of the model is named as such as it describes the data generation process and therefore is the most natural starting point to make modeling assumptions. After assuming the descriptive model the inferential model can be derived which facilitates the means to make inferences about the random effects, Θ, and the hyperparameters (β, A).

1.3 Motivating Examples

To demonstrate why maximum likelihood estimation is often a poor method of estimation let's consider a simple non-trivial example (SNoTE) where the prior mean is assumed to be known and equal to 0, $k = 4$ and $V_i = V$ for all $i = 1, \ldots, k$. From Table 1.1 we see that $y_i \overset{ind}{\sim} N(0, V + A)$ and thus the likelihood is

$$f(\mathbf{y} \mid A) \propto (V + A)^{-\frac{k}{2}} \exp\left(-\frac{\sum_{i=1}^{k} y_i^2}{2(V + A)}\right). \tag{1.6}$$

This leads to the maximum likelihood estimate of

$$\hat{A}_{mle} \equiv \max\left(0, \frac{\sum_{i=1}^{k} y_i^2}{k} - V\right). \tag{1.7}$$

Given that

$$\frac{\sum_{i=1}^{k} y_i^2}{V + A} \sim \chi_k^2 \tag{1.8}$$

we see that the probability the MLE results in an estimate of 0 for A is

$$P\left(\frac{\sum_{i=1}^{k} y_i^2}{k} - V \leq 0\right) = P(\chi_k^2 \leq Bk). \tag{1.9}$$

To obtain a numerical example this probability for $k = 4$ and $B = 0.5$ is calculated to be 0.26. Hence, in over 25% of datasets for a simple example with moderate shrinkage the MLE predicts 100% shrinkages, $\hat{B}_{mle} \equiv \frac{1}{1+\hat{A}_{mle}} = 1$ which can have disastrous results when making inferences about the random effects, θ_i. From (1.5) we see that plugging in the value of $\hat{B}_{mle} = 1$ will result in intervals of 0 length with

arbitrarily high confidence. As θ_i is continuous the probability that these intervals will cover the true value of θ_i is 0. This is further demonstrated in Figure 1.1 where data was simulated 100 times under the true model and the aforementioned plug-in estimation procedure was used.

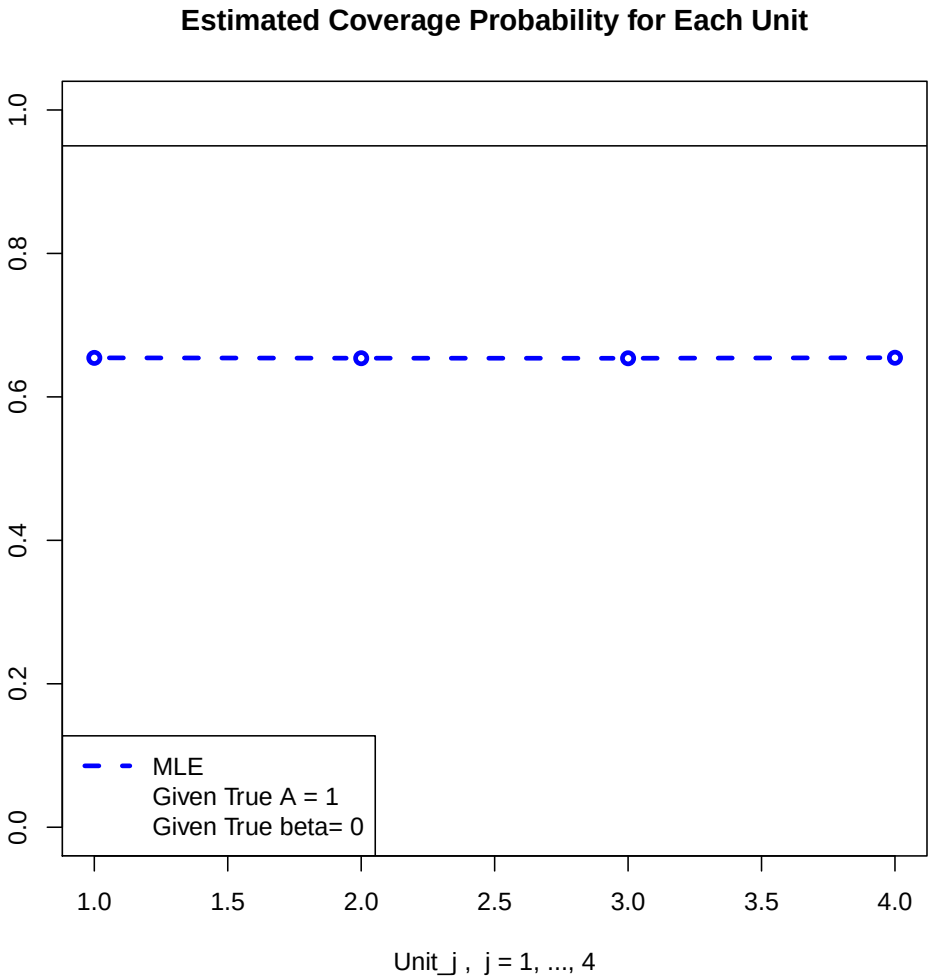

Figure 1.1: Coverage for 95% intervals for MLE when $k = 4$, $V = 1$ and β is known to be 0.

It is noted that in Figure 1.1 the nominal coverage rate is actually closer to 60% rather than the, $(1-0.26)0.95 \approx 70\%$, which might have been expected if the method covered 95% of the time for the datasets where A was not estimated to be 0. This is

largely due to the fact that the plug-in method does not take into account the extra variability present in the estimation of A and assumes that the estimated value is in fact the true value. In Section 1.6.2 we present ways to deal with this problem.

The motivating example demonstrates why MLE and REML based inferences may fail but it is often desirable to see how such methods fare when using real datasets. The dataset which we will refer to here on out as the schools data first appeared in Alderman and Powers (1980) and then in Rubin (1981). This study was conducted by the Education Testing Service (ETS) to test whether student's SAT-V scores are affected by coaching in eight separate schools. The dataset in Table 1.2 contains the estimated coaching effects on SAT-V scores (y_j) and standard errors ($\sqrt{V_j}$) of the eight schools ($j = 1, \ldots, 8$).

Table 1.2: Eight Schools Data

i	y_i	$\sqrt{V_i}$
1	-1	9
2	8	10
3	18	10
4	7	11
5	1	11
6	28	15
7	-3	16
8	12	18

In Figure 1.2 the results of fitting the model with SAS Institute Inc. (2011) are explored. Note that SAS requires that the data be available at the individual student level for each school and not just the group level statistics given in Table 1.2. As such we assume that the variances, V_i, are of the form $V_i = \sigma^2/n_i$ and data was simulated for each school with sample sizes, $n_i = \lfloor \sigma^2/V_i + 0.5 \rfloor$, where $\sigma = 100$. The data was then adjusted such that the means and standard errors agreed with those in Table 1.2. Note that this example is for purely illustrative purposes and does not reflect the analysis had the actual individual school level data been used.

In Figure 1.2 the parameterization and terminology is in the random-fixed effects framework but estimates of the hyperprior parameters from the Normal-Normal model, A and β, can easily be deduced. The estimate of the second level mean, β, is 7.6750 and can be found in the Solution for Fixed Effects table and the estimate of the second level variance, A, is 0 and can be found in the Covariance Parameter Estimates table. To obtain these points estimates SAS Institute Inc. (2011) utilizes restricted maximum likelihood (REML) which considers β a nuisance parameter and constructs a linear combination of the data such that the mean of the transformed data is 0. After transforming, traditional maximum likehood estimation is conducted on the transformed data to estimate the variance component, A. Note that this procedure is equivalent to integrating out β from the likelihood. This allows for the estimation of A whilst still accounting for the loss in degrees of freedom by estimating the unknown prior mean β. It is noted that with this method there is no safeguard against the mode being on the boundary of the parameter space. The asymptotic theory that MLE relies on means that with enough data the mode will eventually

The SAS System

The Mixed Procedure

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	group	0	.	.	.
Residual		9961.57	574.18	17.35	<.0001

Fit Statistics	
-2 Res Log Likelihood	7257.1
AIC (smaller is better)	7259.1
AICC (smaller is better)	7259.1
BIC (smaller is better)	7259.2

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	7.6750	4.0645	7	1.89	0.1009

Solution for Random Effects									
Effect	**Subject**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
Intercept	1	0	.	.	.	.	.	.	.
Intercept	2	0	.	.	.	.	.	.	.
Intercept	3	0	.	.	.	.	.	.	.
Intercept	4	0	.	.	.	.	.	.	.
Intercept	5	0	.	.	.	.	.	.	.
Intercept	6	0	.	.	.	.	.	.	.
Intercept	7	0	.	.	.	.	.	.	.
Intercept	8	0	.	.	.	.	.	.	.

Figure 1.2: Analysis of the eight schools data with SAS

move away from the boundary when normality is achieved. However, for any finite sample there is no guarantee and as such when the mode does occur on the boundary

it suggests that MLE and REML will provide poor estimates and the benefits of the asymptotic theory cannot be used. In this example we see that this is the case as $\hat{A} = 0$. It is noted that SAS also fails to provide a standard error for $\hat{A}$ and fails to construct confidence intervals for the random effects (reparameterized θ_i's). This suggests that $\hat{B}_i = 1$ for all i and that each groups mean should be estimated by $\hat{\beta}$.

In Figure 1.3 we show the average coverage for the intervals of the random effects, θ_i, based on the plug-in estimate procedure for MLE and REML. The coverages are averaged over 100 datasets where the data was generated with the variances, V_i, equal to the schools variances in Table 1.2. The true values of A and β values were set to 117.71 and 8.17 respectively these are the estimates given for the schools data from our proposed procedure noted in Section 1.6. The idea being that the simulated data comes from a data generation process that would likely produce data similar to the schools data in Table 1.2.

Figures 1.1, 1.2 and 1.3 demonstrate that for these examples presented the current methodology based on MLE and REML perform poorly in terms of constructing intervals with good frequency properties for the random effects, θ_i. As such we present a procedure named Gaussian Regression Interactive Multilevel Modeling (GRIMM) that utilizes Bayesian machinery to help overcome some of the problems associated with MLE and REML. The goal of GRIMM is to provide a procedure that results in credible intervals for the random effects that when interpreted as confidence intervals offer the reported nominal coverage. As we are using Bayesian machinery to construct the intervals the first step in constructing our model and our estimation procedure

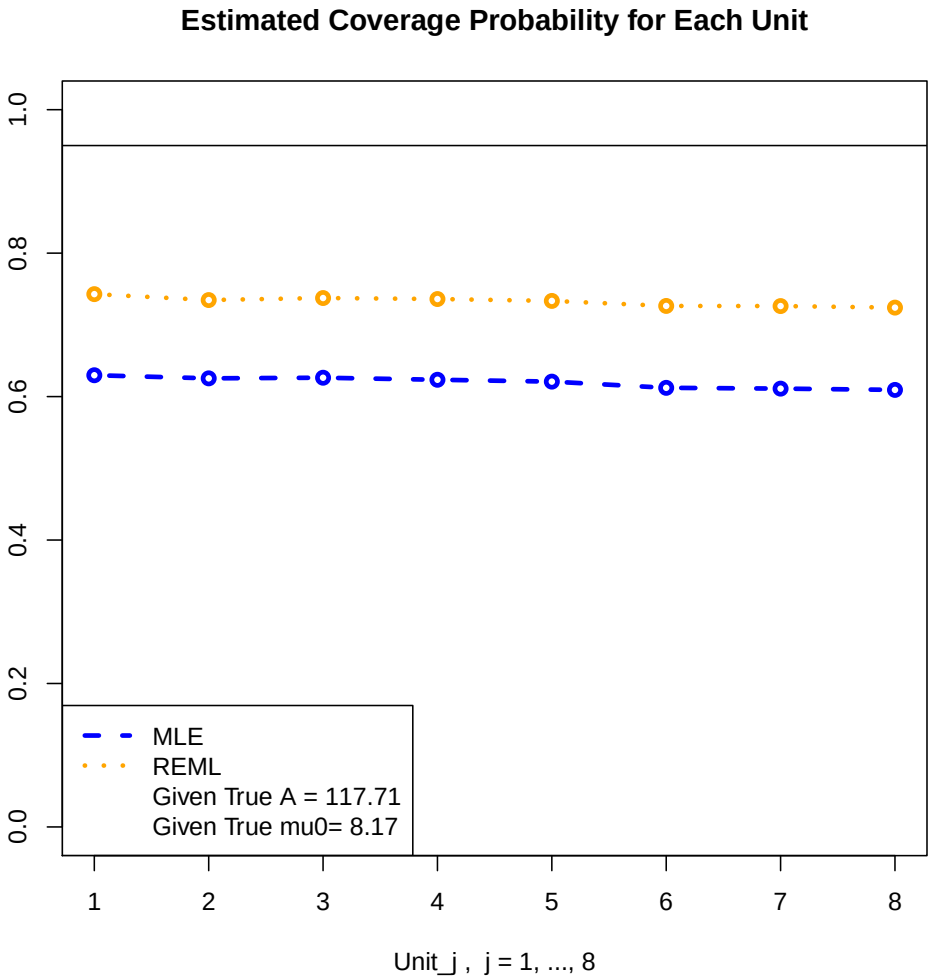

Figure 1.3: Estimated coverages for the eight schools data

is to choose appropriate priors that lead to good frequency properties, as noted in Section 1.4.

1.4 Priors, Posteriors and Propriety

As mentioned in Section 1.1 GRIMM is a procedure that utilizes Bayesian machinery to provide estimates that have good frequency properties and as such we will need to assume reasonable priors that facilitate this need.

Following the suggestions in Morris and Lysy (2012) we assume a flat improper prior on both β and on A.

$$p(\beta, A) \propto 1. \tag{1.10}$$

It is noted that in that in equal variance setting, $V_i = V$ for all i, Morris and Lysy (2012) names this prior Stein's harmonic prior (SHP). They note that the prior is scale invariant and conjugate and that it leads to a formal, Bayes procedure that leads to an admissible and minimax estimator of the shrinkage factor. Further justification of Stein's harmonic prior can be seen due to the fact that the prior $A \sim \text{Unif}(-V, \infty)$, leads to the James-Stein estimator. Knowing that $A > 0$ Stein's harmonic prior seems much more reasonable. Additionally in Morris (1983b), it is noted that under SHP that Stein's unbiased estimate of risk (SURE) is less than the sum of the posterior variances of the random effects indicating that assuming SHP will lead to a conservative procedure. Even though we are considering the unequal variance case it is suggested that this prior is still reasonable and the frequency properties resulting from such a choice of prior is examined in Section 1.8.

Upon assuming the prior noted in (1.10) we can derive the joint posterior distribution

$$
\begin{aligned}
p(\beta, A \mid \mathbf{y}) &\propto \prod_{i=1}^{k} p(y_i \mid \beta, A) \times p(\beta, A) && (1.11)\\
&\propto \prod_{i=1}^{k} N(y_i | x_i^{'}\beta, V_i + A) && (1.12)\\
&= \prod_{i=1}^{k} (V_i + A)^{-\frac{1}{2}} \times \exp\left(\sum_{i=1}^{k} \frac{-(y_i - x_i^{'}\beta)^2}{2(V_i + A)}\right). && (1.13)
\end{aligned}
$$

Note that is often useful to work in a multivariate setting with matrix notation and the marginal distribution of the data can be expressed as a multivariate Normal

$$\mathbf{y} \sim N_k(X\beta, D_{\mathbf{V}+A}) \tag{1.14}$$

where X is the $k \times r$ covariate matrix and $D_{\mathbf{V}+A} \equiv \text{diag}(V_i + A)$. Thus, (1.13) can also be expressed in matrix notation and the two notations will be used interchangeably when appropriate

$$p(\beta, A \mid \mathbf{y}) \propto |D_{\mathbf{V}+A}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - X\beta)^{'} D_{\mathbf{V}+A}^{-1}(\mathbf{y} - X\beta)\right). \tag{1.15}$$

From the joint posterior we can then derive the following conditional posterior distri-

bution for β

$$\begin{aligned} p(\beta|\mathbf{y}, A) &\propto \exp\left(-\frac{1}{2}(\mathbf{y} - X\beta)' D^{-1}_{\mathbf{V}+A}(\mathbf{y} - X\beta)\right) \\ &= \exp\left(-\frac{1}{2}((\mathbf{y} - X\hat{\beta}_A) + X(\hat{\beta}_A - \beta))' D^{-1}_{\mathbf{V}+A}\right. \\ &\left.\times ((\mathbf{y} - X\hat{\beta}_A) + X(\hat{\beta}_A - \beta))\right) \\ &\propto \exp\left(-\frac{1}{2}(\beta - \hat{\beta}_A)' \Sigma_A^{-1}(\beta - \hat{\beta}_A)\right) \end{aligned} \tag{1.16}$$

where $\Sigma_A \equiv (X' D^{-1}_{\mathbf{V}+A} X)^{-1}$ and $\hat{\beta}_A$ is the weighted least squares estimator for β, $\hat{\beta}_A \equiv \Sigma_A X' D^{-1}_{\mathbf{V}+A}\mathbf{y}$. Hence,

$$\beta \mid \mathbf{y}, A \sim N_k(\hat{\beta}_A, \Sigma_A). \tag{1.17}$$

We can also derive the marginal posterior distribution for A, $A \mid \mathbf{y}$

$$\begin{aligned} p(A \mid \mathbf{y}) &= \int_\beta p(\beta, A \mid \mathbf{y}) d\beta \\ &\propto \int_\beta |D_{\mathbf{V}+A}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - X\beta)' D^{-1}_{\mathbf{V}+A}(\mathbf{y} - X\beta)\right) d\beta \\ &= |D_{\mathbf{V}+A}|^{-\frac{1}{2}} |\Sigma_A|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - X\hat{\beta}_A)' D^{-1}_{\mathbf{V}+A}(\mathbf{y} - X\hat{\beta}_A)\right). \end{aligned} \tag{1.18}$$

As we are assuming improper priors it is necessary to prove that our posterior $p(A, \beta \mid \mathbf{y})$ is proper.

Theorem 1.4.1. *If $y_i \sim N(x_i'\beta, V_i + A)$ for $i = 1 \ldots k$ where x_i' is defined as the i^{th} row of the $k \times r$ covariate matrix, X, with rank$(X) = r$ and without loss of generality*

the variances are ordered

$$V_1 \leq V_2 \leq \ldots \leq V_k, \tag{1.19}$$

then the prior $p(A, \beta) \propto 1$ *leads to posterior propriety when* $k \geq r + 3$.

Proof. To prove posterior propriety we must show that

$$\int_0^\infty \int_\beta p(A, \beta \mid \mathbf{y}) d\beta dA < \infty. \tag{1.20}$$

We then note that

$$\begin{aligned}
\int_0^\infty \int_\beta p(A, \beta \mid \mathbf{y}) d\beta dA &= \int_0^\infty p(A \mid \mathbf{y}) dA \\
&= \int_0^\infty |D_{\mathbf{V}+A}|^{-\frac{1}{2}} \mid X' D_{V_i+A}^{-1} X|^{-\frac{1}{2}} \\
&\times \exp\left(-\frac{1}{2}(\mathbf{y} - X\hat{\beta}_A)' D_{\mathbf{V}+A}^{-1}(\mathbf{y} - X\hat{\beta}_A)\right) dA \\
&\leq \int_0^\infty |D_{\mathbf{V}+A}|^{-\frac{1}{2}} |X' D_{\mathbf{V}+A}^{-1} X|^{-\frac{1}{2}} dA
\end{aligned} \tag{1.21}$$

where $D_{V_k+A} \equiv (V_k + A) I_k$.

It is noted that

$$D_{\mathbf{V}+A}^{-1} \geq D_{V_k+A}^{-1} \tag{1.22}$$

and therefore

$$\begin{aligned}
|X' D_{\mathbf{V}+A}^{-1} X| &\geq \mid X' D_{V_k+A}^{-1} X| \\
&= (V_k + A)^{-r} |X' X|.
\end{aligned} \tag{1.23}$$

Similarly we see that

$$\begin{aligned} |D_{\mathbf{V}+A}|^{-\frac{1}{2}} &= \prod_{i=1}^{k}(V_i + A)^{-\frac{1}{2}} \\ &\leq (V_1 + A)^{-\frac{k}{2}}. \end{aligned} \tag{1.24}$$

Therefore

$$\begin{aligned} \int_0^\infty \int_\beta p(A, \beta \mid \mathbf{y}) d\beta dA &\leq \int_0^\infty (V_1 + A)^{-\frac{k}{2}} (V_k + A)^{\frac{r}{2}} |X^{'} X|^{-\frac{1}{2}} dA \\ &\propto \int_0^\infty (V_1 + A)^{-\frac{k-r}{2}} \left(\frac{V_k + A}{V_1 + A}\right)^{\frac{r}{2}} dA \\ &\leq \left(\frac{V_k}{V_1}\right)^{\frac{r}{2}} \int_0^\infty (V_1 + A)^{-\frac{k-r}{2}} dA \\ &= \left(\frac{V_k}{V_1}\right)^{\frac{r}{2}} \left[\frac{-2(V_1 + A)^{-\frac{k-r-2}{2}}}{k - r - 2}\right]_0^\infty \end{aligned} \tag{1.25}$$

which is finite for $k - r \geq 3$.

□

Intuitively the requirement for $k - r \geq 3$ is due to the fact that conceptually when covariates are present shrinkage to 0 can be thought to occur on the residuals of the regression where r degrees of freedom has been lost due to the estimation of the regression coefficient β.

Although we present our own proof for Theorem 1.4.1 more general results can be found in Berger et al. (2005) and Michalak and Morris (2014). It is also noted that Morris and Lysy (2012) examined posterior propriety for a class of priors for the canonical equal variances case ($V_i = V$, for all i) where no covariates are present and the prior mean is assumed to be 0. The class of priors was characterized by a three

parameter family of distributions constructed by combining the densities of the:

1. Scale-invariant priors on A. Indexed by $c \geq 0$ and of the form

$$p(A)dA \propto A^{c/2}\frac{dA}{A} \text{ where } A > 0. \tag{1.26}$$

As a distribution on B this corresponds to

$$p(B)dB \propto B^{-c/2-1}(1-B)^{c/2-1}dB \text{ where } 0 < B = \frac{V}{V+A} < 1. \tag{1.27}$$

2. Conjugate priors on B. Indexed by $k_0 > 0$ and $S_0 \geq 0$ and of the form

$$p(B)dB \propto B^{(k_0-2)/2}\exp(-BS_0/2)dB/B \text{ where } 0 < B < 1. \tag{1.28}$$

The resultant prior on B being,

$$p(B \mid k_0, c, S_0) \propto B^{(k_0-c)/2-1}(1-B)^{c/2-1}\exp(-BS_0/2). \tag{1.29}$$

Theorem 1.4.2. *For the case of k groups, r covariates, unequal variances and a flat prior on β the three-parameter family of priors in* (1.29) *on, $B_1 = \frac{V_1}{V_1+A}$, of the form where $0 < B_1 < 1$ and $V_1 = min(V_i)$ lead to posterior propriety conditional on $S_0 \geq 0$ and $k_0 > u - k^*$ where $u = k^* + k_0 - c$ and $k^* = k - r$.*

Proof. As in Morris and Lysy (2012) we will set $S_0 = 0$ this is because $\exp(-B_1S_0/2)$ is always bounded by 1 for $S_0 \geq 0$ and so the value of S_0 is inconsequential in our

proof of posterior propriety as we can always bound our posterior from above. This now leads our class of priors to have the form

$$B_1 \sim \text{Beta}(\frac{1}{2}(u-k^*), \frac{1}{2}(k_0-(u-k^*))). \tag{1.30}$$

From the proof of Theorem 1.4.1 we note that after integrating out β we can bound the likelihood, $p(\mathbf{y} \mid A)$, up to a proportionality constant by

$$(V_1+A)^{-\frac{k-r}{2}} \propto B_1^{\frac{k*}{2}} \tag{1.31}$$

hence leading to the fact that

$$\begin{aligned}\int_0^1 p(B_1 \mid \mathbf{y})dB_1 &\leq \text{constant} \times \int_0^1 B_1^{\frac{k*}{2}} \times B_1^{\frac{u-k^*}{2}-1}(1-B_1)^{\frac{k^*+k_0-u}{2}-1}dB_1 \\ &\propto \int_0^1 B_1^{\frac{u}{2}-1}(1-B_1)^{\frac{k_0-(u-k^*)}{2}-1}dB_1.\end{aligned} \tag{1.32}$$

As (1.32) is of the form of a Beta density we can see that this integral will be finite under the condition that $u > 0$ and $k_0 > u - k^*$ or equivalently $c \geq 0$ and hence posterior propriety is proved for these conditions. □

The class of priors described in Theorem 1.4.2 do have some interesting cases as noted in Figure 1.4.

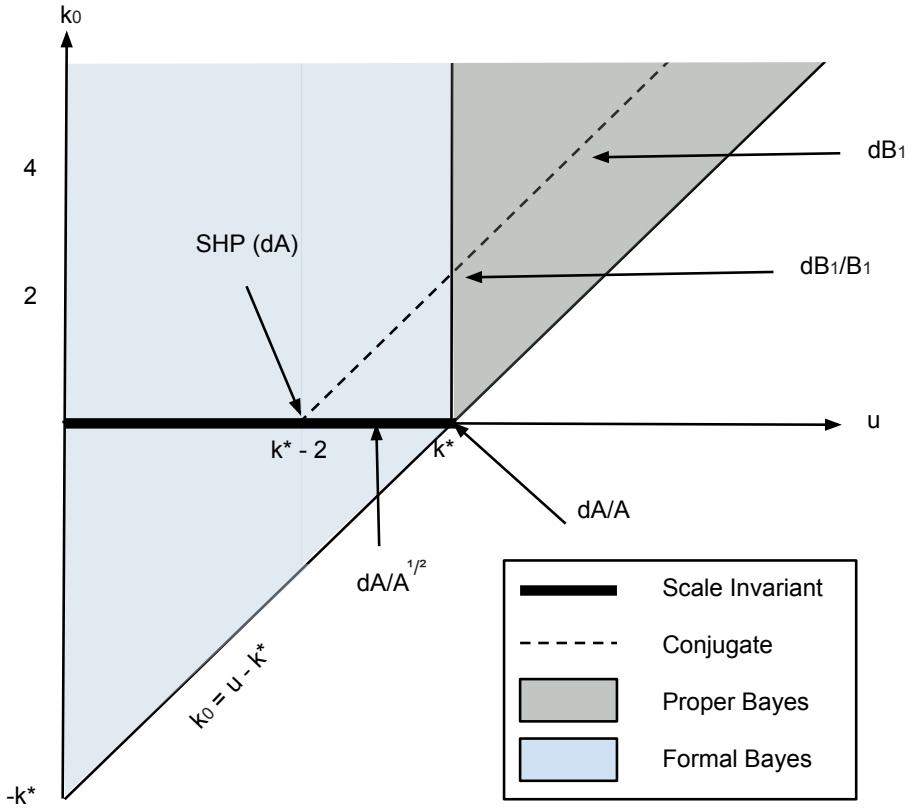

Figure 1.4: Properties of the class of priors described in Theorem 1.4.2

In particular note that:

1. The 45° line, $k_0 = u - k^*$, marks the divide between proper (above) and improper (below) posteriors.

2. As per the condition $u > 0$ we are restricted to positive values of u.

3. Proper priors lead to proper Bayes procedures and require that $u > k^*$ or equivalently $k_0 > c$.

4. Scale-invariant priors require $k_0 = 0$ and $0 < u < k^*$ or equivalently $c \geq 0$.

5. Conjugate priors lie on the line $k_0 = u - k^* - 2$ thus eliminating the factor of $(1 - B_1)$ from the prior distribution.

6. As noted in Morris and Lysy (2012) Jeffrey's prior dA/A should not be used due to leading to 100% shrinkage in the equal variance case.

7. Stein's harmonic prior (dA) seems like a natural choice as it is invariant, conjugate and leads to a proper posterior.

1.5 Equal Variances

In the equal variance setting $V_i = V$ for all $i = 1 \dots k$ a symmetry arises in the problem making inferences a lot more tractable. We see that (1.17) reduces to

$$\beta \mid \hat{\beta}, A \sim N(\hat{\beta}, (V + A)(X'X)^{-1}) \tag{1.33}$$

where $\hat{\beta} \equiv (X'X)^{-1}X'\mathbf{y}$ is the traditional least squares estimator and (1.18) reduces to

$$p(A \mid S) \propto (V + A)^{-\frac{k-r}{2}} e^{-\frac{S}{2V}} \tag{1.34}$$

where $S \equiv \sum_{i=1}^{k} (y_i - x_i'\hat{\beta})^2$. The distribution of the random effects is also simpler due to the fact that each group, i, now shares a common shrinkage factor $B \equiv V/(V + A)$

$$\theta_i \mid A, \beta \sim N((1 - B)y_i + Bx_i'\beta, V(1 - B)). \tag{1.35}$$

1.5.1 Point Estimation

A primary goal of GRIMM is to make inference about the random effects, θ_i. As seen in (1.35) the mean and variance of θ_i are linear in the shrinkage factor B. A secondary goal is thus to estimate B with an estimator such that $\hat{B} \equiv E(B \mid \mathbf{y})$.

To obtain point estimates for the shrinkage factor, B, we can perform a change of

variable for the density stated in (1.34) and derive the posterior distribution

$$p(B \mid S) \propto (V+A)^{-\frac{k-r}{2}} e^{-\frac{S}{2(V+A)}} \times (V+A)^2 \tag{1.36}$$

$$\propto B^{\frac{k-r-2}{2}-1} e^{-\frac{SB}{2V}} \tag{1.37}$$

and from (1.37) we note that

$$B \mid S \sim \text{Gamma}\left(\frac{k-r-2}{2}, \frac{S}{2V}\right) \text{ where } 0 < B < 1. \tag{1.38}$$

Given (1.38) and letting $a \equiv \frac{k-r-2}{2}$ and $b \equiv \frac{S}{2V}$ it can be seen that the expectation of any power, p, of B is

$$E(B^p \mid S) = \frac{\int_0^1 B^p b^a \Gamma(a)^{-1} B^{a-1} e^{-bB} dB}{P(G_{a,b} < 1)} \tag{1.39}$$

$$= b^{-p} \frac{\Gamma(a+p)}{\Gamma(a)} \frac{P(G_{a+p,b} < 1)}{P(G_{a,b} < 1)}. \tag{1.40}$$

It is noted that this is an extension from the case in Morris and Lysy (2012) to the expectation of any power of B when covariates are present. From (1.40) we see that a reasonable point estimate for B is

$$\hat{B} \equiv E(B \mid S) = \frac{(k-r-2)V}{S} \times \frac{P(\chi^2_{k-r} < S/V)}{P(\chi^2_{k-r-2} < S/V)}. \tag{1.41}$$

and its variance

$$v \equiv \text{Var}(B|S) = E(B^2 \mid S) - (E(B|S))^2 \tag{1.42}$$

where from 1.40 $E(B^2 \mid S)$ is given by

$$E(B^2 \mid S) = (k-r-2)(k-r) \times \frac{V^2}{S^2} \times \frac{P(\chi^2_{k-r+2} < S/V)}{P(\chi^2_{k-r-2} < S/V)}. \tag{1.43}$$

1.5.2 Direct Posterior Sampling

While point estimates have the benefit of being quickly computed sometimes inferences require knowledge about the entire posterior distribution. Luckily, due to the symmetry in the problem samples can easily be drawn from the full joint posterior by the direct posterior sampling algorithm noted in Algorithm 1.1.

Algorithm 1.1 Direct posterior sampling for equal variances

for t in $1:T$ **do**
  $B^{(t)} \leftarrow 1$
  while $B^{(t)} \geq 1$ **do**
    $B^{(t)} \leftarrow$ draw from $B \mid S$ in (1.38)
  end while
  $A^{(t)} \leftarrow V(1-B^{(t)})/B^{(t)}$
  $\beta^{(t)} \leftarrow$ draw from $\beta \mid \hat{\beta}, A^{(t)}$
  for i in $1:k$ **do**
    $\theta_i^{(t)} \leftarrow$ draw from $\theta_i \mid \beta^{(t)}, A^{(t)}, y_i$ (1.5)
  end for
end for
Return: $A^{(1:T)}$, $\beta^{(1:T)}$, $\Theta^{(1:T)}$

1.5.3 Example: Hospitals

In a 1992 study of medical profiling evaluation of 23 New York hospitals the estimated successful outcome rates for patients following coronary artery bypass graft (CABG) was obtained. (Morris and Lysy (2012)).

Table 1.3: New York hospital data with unequal variances

i	y_i	$\sqrt{V}_i$	d_i	n	i	y_i	$\sqrt{V}_i$	d_i	n_i
1	-0.14	1.22	10	347	13	-0.08	0.96	16	563
2	-1.21	1.22	13	349	14	0.61	0.93	14	593
3	-1.43	1.20	14	358	15	2.05	0.93	9	602
4	1.56	1.14	7	396	16	0.57	0.91	15	629
5	0.00	1.10	12	431	17	1.10	0.90	13	636
6	0.41	1.08	11	441	18	-2.42	0.84	35	729
7	0.08	1.04	13	477	19	-0.38	0.78	26	849
8	-2.15	1.03	22	484	20	0.07	0.75	25	914
9	-0.34	1.02	15	494	21	0.96	0.74	20	940
10	0.86	1.02	11	501	22	-0.21	0.66	35	1193
11	0.01	1.01	14	505	23	1.14	0.62	27	1340
12	1.11	0.98	11	540					

In Table 1.3 n_i represents the number of trials for hospital i and d_i represents the number of successful surgeries. Naturally this data is most suitable for a Beta-Binomial hierarchical model and in fact this data can easily be fit using a procedure known as Binomial Regression Interactive Multilevel Modeling (BRIMM) using the software provided in Kelly et al. (2014a) and described in Kelly et al. (2014b). To investigate the Normal-Normal model Morris and Lysy (2012) make an appropriate variance stabilizing arcsin transformation on the estimated success rate estimates

d_i/n_i to obtain y_i and $\sqrt{V_i}$.

For illustrative purposes a subset of hospitals (8 to 15) whose variances were closest to 1 were chosen such that an equal variance model can be fitted. The equal variance Normal-Normal model is a lot more tractable than the unequal variance case and so it is useful to have an illustrative dataset such as the one listed in Table 1.4. Analysis of the full hospital data is presented in Section 1.7 utilizing the methods for unequal variances presented in Section 1.6.

Table 1.4: New York hospital data with equal variances

	y_i	$\sqrt{V_i}$
1	-2.15	1.00
2	-0.34	1.00
3	0.86	1.00
4	0.01	1.00
5	1.11	1.00
6	-0.08	1.00
7	0.61	1.00
8	2.05	1.00

We can now apply the direct posterior sampling technique noted in Algorithm 1.1 to the equal variance data presented in Table 1.4. The results are presented in Table 1.5 where s_i is defined as the estimate of the posterior standard deviation and $(\hat{\theta}_{i,\alpha/2}, \hat{\theta}_{i,1-\alpha/2})$ is the $(1-\alpha)\%$ confidence interval.

Table 1.5: Direct posterior sampling results for the equal variances hospital data

	y_i	$\sqrt{V_i}$	$\hat{\beta}$	$\hat{B}_i$	$\hat{\theta}_{i,0.025}$	$\hat{\theta}_i$	$\hat{\theta}_{i,0.975}$	s_i
1	-2.15	1.00	0.26	0.42	-3.14	-1.14	0.57	0.97
2	-0.34	1.00	0.26	0.42	-1.75	-0.09	1.45	0.81
3	0.86	1.00	0.26	0.42	-0.93	0.61	2.27	0.81
4	0.01	1.00	0.26	0.42	-1.50	0.11	1.68	0.80
5	1.11	1.00	0.26	0.42	-0.78	0.75	2.45	0.82
6	-0.08	1.00	0.26	0.42	-1.56	0.06	1.62	0.80
7	0.61	1.00	0.26	0.42	-1.09	0.46	2.09	0.80
8	2.05	1.00	0.26	0.42	-0.29	1.30	3.17	0.89

It is worth noting the following in Table 1.5. The shrinkage value of 0.42 represents a considerable amount of shrinkage and that the posterior mean estimates have been pulled toward the estimated prior mean of 0.26. The posterior standard deviations of the random effects are all smaller than the observed standard deviation of 1. This is due to the fact that we are borrowing information between the hospitals to improve upon the estimate of each hospital's mean over the observed value. It's also noted that the ordering of the hospitals with respect to how successful they are at performing CABG surgeries has not changed. Later when we tackle the unequal variance problem in Section 1.7 we may see the ordering of the hospitals change due to varying amounts of information present within each hospital given that each hospital would have performed a different number of surgeries. Orderings can also change if covariates are present.

1.6 Unequal Variances

In the case of unequal variances the lack of symmetry in the problem prevents the use of exact inferences noted in Section 1.5. In this section we examine how maximum likelihood (MLE) and restricted maximum likelihood (REML) inferences compare with a procedure suggested in Morris and Tang (2011) known as adjustment for density maximization (ADM).

1.6.1 Adjustment for Density Maximization

Morris (1988) suggests improvements to approximating univariate probability densities based on only calculating two derivatives. The method adjusts a density by multiplying it by a factor such that the mode of the adjusted density is close to the mean. This is accomplished by approximating a density from one chosen from the Pearson family with the adjustment factor being determined by the choice of the approximating distribution.

As an example let's suppose we have the posterior distribution $\lambda \mid x \sim \text{Expo}(x)$ so that $E(\lambda|x) = 1/x$. We would like to construct a point estimate of λ, $\hat{\lambda}$, using differentiation such that $\hat{\lambda} = E(\lambda \mid x)$. Note that a maximum a posteriori (MAP) estimate would give $\hat{\lambda}_{MAP} = 0$ but by maximizing an adjusted density, $\lambda \times p(\lambda|x)$,

the maximum is obtained such that $\hat{\lambda} = E(\lambda \mid x)$

$$\begin{aligned} \text{Adjusted density} &= \lambda \times p(\lambda \mid x) \\ &= \lambda x e^{-\lambda x} \end{aligned}$$

and therefore $\hat{\lambda} = E(\lambda|x) = 1/x$. The adjustment factor, λ, comes from the fact that the Poisson is conjugate to the Exponential distribution and the quadratic variance function of a Poisson(λ) is λ. This is demonstrated for an empirical example in Figure 1.5.

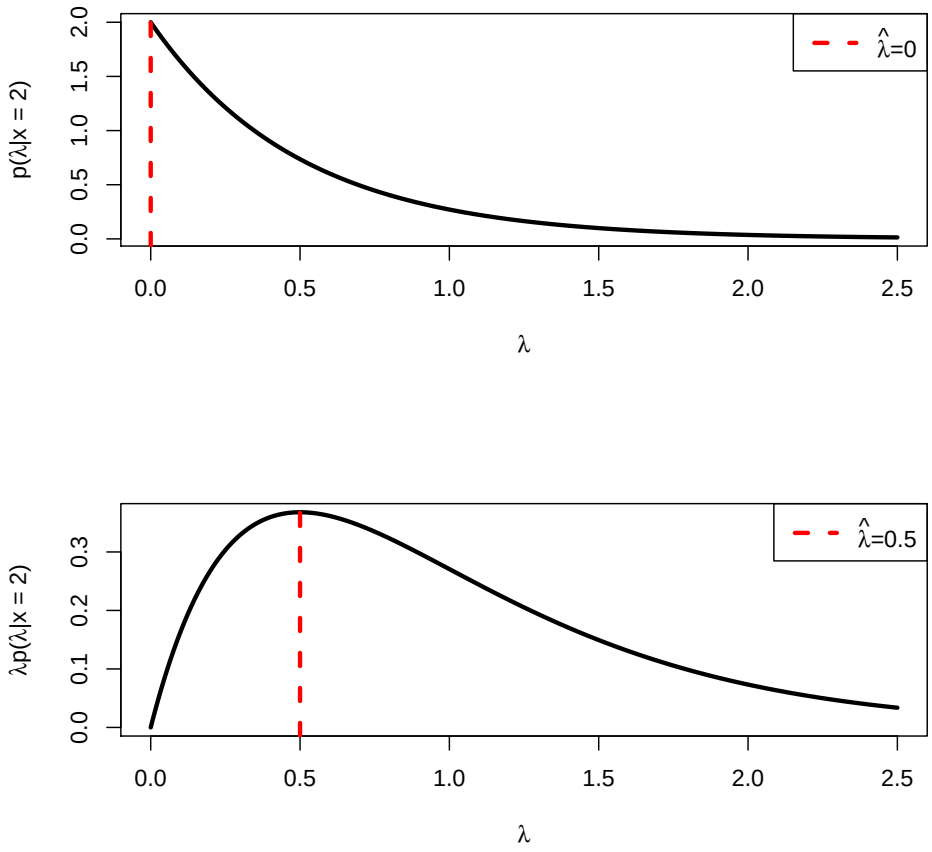

Figure 1.5: ADM empirical Exponential example where $x = 2$.

In the Normal-Normal hierarchical model our quantity of interest, θ_i, has mean and variance linear in the shrinkage factor, B_i, as demonstrated by (1.5). As shrinkage factors, $0 < B_i < 1$, and we wish to obtain estimates such that $\hat{B}_i = E(B_i \mid \mathbf{y})$ an ADM approximation with Beta as an approximating distribution seems reasonable. Let's consider the Beta(a, b) distribution noted in Figure 1.6 with density

$$p(B) \propto B^{a-1}(1-B)^{b-1}. \tag{1.44}$$

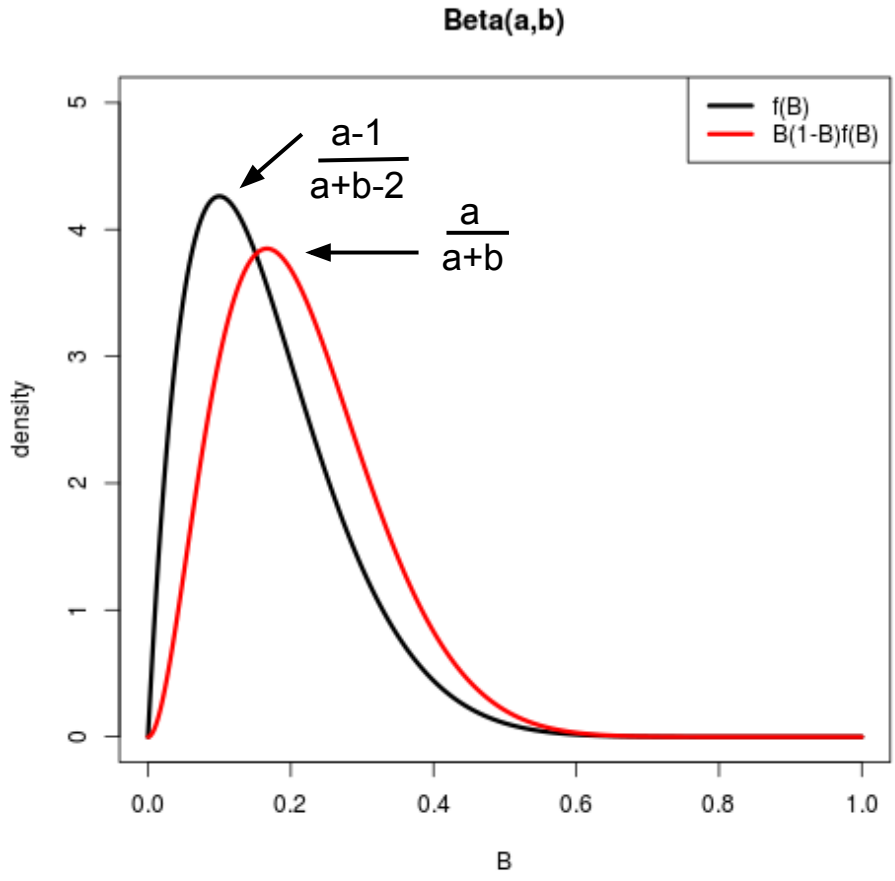

Figure 1.6: ADM Beta example

The mode of such a distribution is $\frac{a-1}{a+b-2}$. However, by multiplying the density $f(B)$ by $B(1-B)$ the mode of the adjusted density is $\frac{a}{a+b}$, the mean of the original Beta distribution. It's clear that if the original density is the same as the approximating one then this procedure will be exact, however, if the Pearson family is only approximate then the procedure will only give an approximation to the mean. In Morris and Tang (2011) it is shown that this adjustment is equivalent to multiplying $p(\mathbf{y}|A)$ by A. In Morris (1988) it is noted that the adjustment factor is the quadratic variance function (as a function of the mean) of the distribution of the natural exponential family to which the distribution from the Pearson family is conjugate. In Table 1.6 we note the results found in Morris (1988) and list the approximating distributions from the Pearson family and the relevant adjustment factors, $V(x)$, for each under the ADM framework.

Table 1.6: Approximating distributions for ADM and their adjustment factors $V(x)$. In this table $a, b > 0$ and $n > 1$.

Approx. Dist.	**Density p(x)** $\propto$	**Support**	**Conjugate to Dist.**	**V(x)**
Normal(μ, σ^2)	$e^{-(x-\mu)^2/2\sigma^2}$	$(-\infty, \infty)$	Normal	1
Gamma(a, b)	$x^a e^{-bx}$	$(0, \infty)$	Poisson	x
Inv-Gam(a,b)	$x^{-a-1}e^{-b/x}$	$(0, \infty)$	Gamma	x^2
Beta(a, b)	$x^{a-1}(1-x)^{b-1}$	$(0, 1)$	Binomial	$x(1-x)$
$F^*(a, b)$	$\frac{x^a}{(1+x)^{a+b-1}}$	$(0, \infty)$	Negative Binomial	$x(1+x)$
t_n	$(1+\frac{x^2}{n})^{-\frac{n-1}{2}}$	$(-\infty, \infty)$	NEF-CHS	$n+x^2$

In a frequentist setting, it is noted that the method can be thought of as a more

general procedure of maximum likelihood estimation where instead of fitting the two derivatives to a Normal distribution a distribution is chosen from one of the Pearson families noted in Table 1.6. The choice of the Normal distribution in maximum likelihood estimation is only a good approximation when the true distribution is Normal or there is a large sample size. For small sample sizes there may be skewness present and a bounded or semi-bounded distribution may be more appropriate. As the Normal distribution is a member of the Pearson family the procedure is consistent with maximum likelihood estimation. From the Bayesian perspective this procedure can be thought of as approximating posterior means and variances via differentiation instead of integration.

To see how ADM applies in the context of the Normal-Normal hierarchical we can first construct the marginal log-likelihood from (1.18)

$$\begin{aligned} l(A) &= \frac{1}{2}\log(|\Sigma_A|) + \sum_{i=1}^{k} \log(N(y_i \mid x_i^{'}\hat{\beta}_A, V_i + A)) \\ &= \frac{1}{2}\log(|\Sigma_A|) - \frac{1}{2}\sum_{i=1}^{k} \log(V_i + A) - \sum_{i=1}^{k} \frac{(y_i - x_i^{'}\hat{\beta}_A)^2}{2(V_i + A)} \end{aligned} \tag{1.45}$$

The adjusted log-likelihood is thus

$$l_{ADM}(A) \equiv \log(A) + l(A) \tag{1.46}$$

and due to the fact that our prior assumes, $P(\beta, A) \propto 1$, it is noted that the adjusted log posterior is identical to the adjusted log-likelihood. As no closed form solutions exist to maximize this quantity methods such as Newton-Raphson or the EM-based

procedures noted in Section 2.3 can be used. Therefore,

$$\hat{A} \equiv \arg\max_{A} l_{ADM}(A) \tag{1.47}$$

where the first and second derivatives of (1.46) to be used in a Newton-Raphson procedure are noted in Tang (2002). In Morris and Tang (2011) it is noted that the distribution for each shrinkage factor, B_i, can be approximated via the ADM procedure with a Beta distribution with mean and variance approximated by

$$E(B_i \mid \mathbf{y}) \approx \hat{B}_i = \frac{V_i}{V_i + \hat{A}} \tag{1.48}$$

$$Var(B_i \mid \mathbf{y}) \approx v_i = \frac{\hat{B}_i^2(1-\hat{B}_i)^2}{-\frac{\partial^2 l_{ADM}}{\partial \alpha^2}|_{\alpha=\hat{\alpha}} + \hat{B}_i(1-\hat{B}_i)} \tag{1.49}$$

where $\alpha = \log(A)$. This leads to the corresponding approximating Beta(a_i, b_i) distribution where

$$a_i \equiv \frac{-\frac{\partial^2 l_{ADM}}{\partial \alpha^2}|_{\alpha=\hat{\alpha}}}{1-\hat{B}_i} \tag{1.50}$$

$$b_i \equiv \frac{-\frac{\partial^2 l_{ADM}}{\partial \alpha^2}|_{\alpha=\hat{\alpha}}}{\hat{B}_i} \tag{1.51}$$

To demonstrate the dangers of maximum likelihood estimation and why ADM might be more appropriate let's consider the schools data presented in Section 1.3.

In Figure 1.7 we can compare the likelihood and adjusted likehood after integrating out β as in Section 1.4 (equivalent to the REML procedure). Here we see that the mode of the distribution, even after undertaking the REML adjustment, lies on the

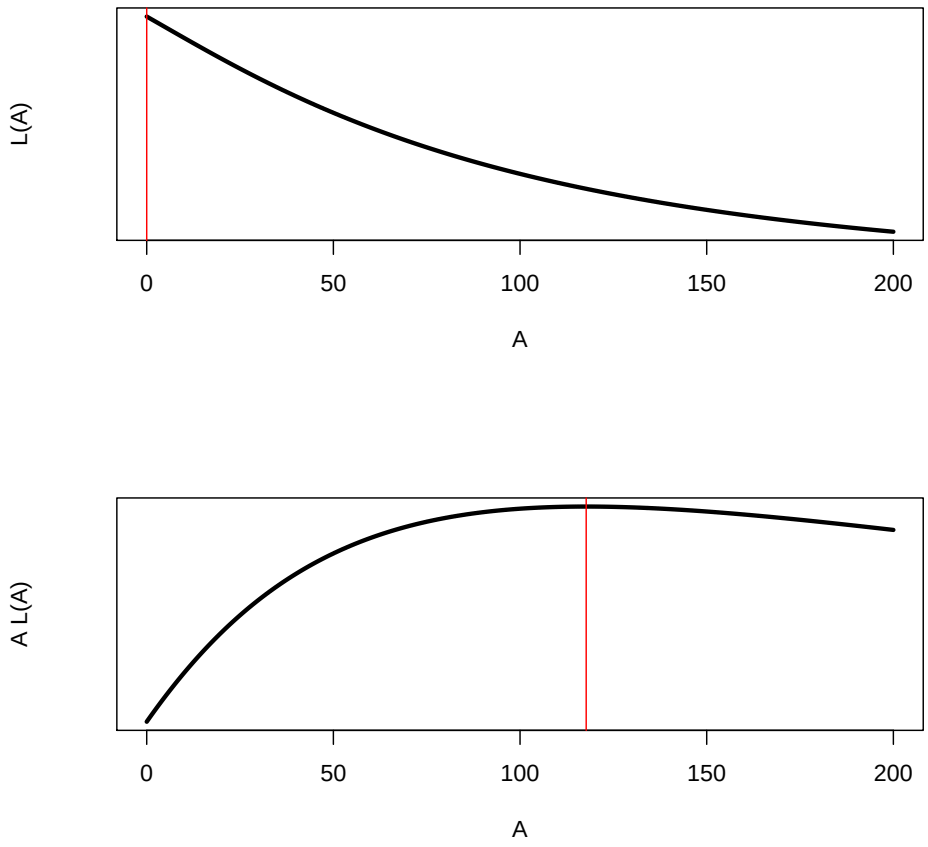

Figure 1.7: Likelihood and adjusted likelihood for the eight schools data

boundary of the parameter space which when occurs violates an assumption of MLE and REML estimation. In this example we see that A would be estimated, $\hat{A} = 0$ suggesting that $B_i = 1$ for all i and that each groups mean should be estimated by $\hat{\beta}$.

1.6.2 Inferences for θ_i

As mentioned in Section 1.6.1 the mean and variances of the random effects, θ_i, are linear in the shrinkage factors, B_i and the ADM procedure provides us with an estimation method that aims to estimate $E(B_i \mid \mathbf{y})$. Although shrinkage factors

are interesting in the their own right the primary goal of analysis for our purposes is to make inferences about the random effects conditional on the data, $\theta_i \mid \mathbf{y}$, for $i = 1 \ldots k$. To approximate the mean and variances of the random effects we first note that

$$\theta_i | y_i, \beta, A \overset{ind}{\sim} N((1 - B_i)y_i + B_i x_i^{'}\beta, V_i(1 - B_i)) \text{ for } i = 1 \ldots k. \tag{1.52}$$

The mean and variances of θ_i can therefore be calculated using the affectionately known Adam's and EVE's laws. The first step of Adam's and EVE's law will be presented here as we will utilize this result in Section 2.2.

$$\begin{aligned} E(\theta_i \mid \mathbf{y}, A) &= E(E(\theta_i \mid y_i, A, \beta) \mid y_i, A) \\ &= (1 - B_i)y_i + B_i x_i^{'}\hat{\beta}_A. \end{aligned} \tag{1.53}$$

$$\begin{aligned} \text{Var}(\theta_i \mid \mathbf{y}, A) &= E(\text{Var}(\theta_i \mid y_i, A, \beta) \mid y_i, A) \\ &\quad + \text{Var}(E(\theta_i \mid y_i, A, \beta) | y_i, A) \\ &= V_i(1 - B_i) + B_i^2 x_i^{'}\Sigma_A x_i. \end{aligned} \tag{1.54}$$

Applying Adam and Eve's law again (over A) and making approximation via the adjustment for density procedure noted in 1.6.1 is demonstrated in Tang (2002) the final results of which we present below.

$$\hat{\theta}_i \equiv E(\theta_i \mid \mathbf{y}) \approx (1 - \hat{B}_i)y_i + \hat{B}_i x_i^{'}\hat{\beta}_{\hat{A}} \tag{1.55}$$

where $\hat{B}_i \equiv \frac{V_i}{V_i+\hat{A}}$ and $\hat{\beta}_{\hat{A}} \equiv \Sigma_{\hat{A}} X' W_{\hat{A}} \mathbf{y}$ where $\hat{A}$ is obtained using the adjustment for density maximization procedure noted in Section 1.6.1.

The variance of the random effects can be estimated by

$$\begin{aligned} s_i^2 \equiv \text{Var}(\theta_i \mid \mathbf{y}) \approx V_i(1-\hat{B}_i) + (x_i' \Sigma_{\hat{A}} x_i)\hat{B}_i^2 \\ + v_i \left(x_i' \Sigma_{\hat{A}} x_i - (V_i + \hat{A}) x_i' M x_i + (y_i - x_i' \hat{\beta}_{\hat{A}} + (V_i + \hat{A}) x_i' u)^2 \right) \end{aligned} \tag{1.56}$$

where

$$M \equiv \frac{\partial \Sigma_A}{\partial A}|_{A=\hat{A}} \tag{1.57}$$

$$u \equiv \frac{\partial \beta_A}{\partial A}|_{A=\hat{A}}. \tag{1.58}$$

It is noted in Morris and Tang (2011) that a Normal approximation using the approximate means and variances noted in (1.55) and (1.54) respectively can be used to construct confidence intervals for the θ_i's but that there may be skewness present in the distributions of these random effects. In Section 1.6.3 we propose a method to incorporate skewness into our inferences.

1.6.3 Skewness

To improve upon posterior intervals that utilize a Normal approximation we incorporate the fact that the posterior $\theta_i|\mathbf{y}$ may be skewed. Given that

$$\theta_i|y_i, \beta, A \overset{ind}{\sim} N((1-B_i)y_i + B_i x_i^{'}\beta, V_i(1-B_i)) \text{ for } i = 1 \dots k. \tag{1.59}$$

and

$$\beta|\mathbf{y}, A \overset{ind}{\sim} N(\hat{\beta}_A, \Sigma_A) \tag{1.60}$$

and due to the fact that the mean and variance of θ_i are linear in the shrinkage factors, B_i, it seems likely that the skewness present in $\theta_i \mid \mathbf{y}$ is largely due to the skewness present in the shrinkage factors, B_i, and not due to the distribution of β. This was checked for a few examples and it seemed to hold true in the cases we tested. As such, to obtain a reasonable approximation for the skewness we can use the approximation that β is known and equal to the estimated value $\hat{\beta}_{\hat{A}}$ with knowledge that this assumption should have little effect on the skewness approximation. Note that each calculation below is conditional on ($\mathbf{y}$ and assumes that the true value of β is the estimated value the notation of which we have omitted for the sake of readability.

We can calculate the third central moment, μ_3, using the law of third cumulants (Brillinger (1969))

$$K_3 = EK_3 + K_3E + 3\text{Cov}(E, V). \tag{1.61}$$

Applying this law to θ_i and by conditioning on A we see that

$$\mu_3(\theta_i) = E(\mu_3(\theta_i|A)) + \mu_3(E(\theta_i|A)) + 3\text{Cov}(E(\theta_i|A), \text{Var}(\theta_i|A)). \tag{1.62}$$

We can calculate or approximate each term in (1.62)

$$\begin{aligned} \mu_3(E(\theta_i) \mid A) &= \mu_3((1 - B_i)y_i + B_i x_i^{'}\hat{\beta}_{\hat{A}}) \\ &= -(y_i - x_i^{'}\hat{\beta}_{\hat{A}})^3 \mu_3(B_i) \end{aligned} \tag{1.63}$$

$$E(\mu_3(\theta_i|A)) = 0 \text{ from (1.59)} \tag{1.64}$$

and

$$\begin{aligned} \text{Cov}(E(\theta_i|A), \text{Var}(\theta_i|A)) &= \text{Cov}((1 - B_i)y_i + B_i x_i^{'}\beta, V_i(1 - B_i)) \\ &= V_i(y - x_i^{'}\hat{\beta}_{\hat{A}})\text{Var}(B_i). \end{aligned} \tag{1.65}$$

Naturally a reasonable approximation for $\mu_3(B_i)$ and $\text{Var}(B_i)$ would be to utilize the Beta approximation implied by the ADM approximation. Hence, $\text{Var}(B_i)$ can be approximated by v_i as noted in (1.49) and $\mu_3(B_i)$ can be approximated by the third central moment of the corresponding Beta distribution

$$\mu_3(B_i) \approx c_i \equiv \frac{2a_i b_i (b_i - a_i)}{(a_i + b_i)^3 (a_i + b_i + 2)\sqrt{a_i + b_i + 1}} \tag{1.66}$$

where a_i and b_i are given in (1.50) and (1.51) respectively.

Hence we can approximate the third central moment of θ_i as

$$\mu_3(\theta_i \mid \mathbf{y}) = \widehat{\mu_3(\theta_i \mid \mathbf{y})} \approx -(y_i - x_i'\hat{\beta}_{\hat{A}})^3 c_i + 3V_i(y_i - x_i'\hat{\beta}_{\hat{A}})v_i. \tag{1.67}$$

After approximating the first three central moments of $\theta_i \mid \mathbf{y}$ with $\hat{\theta}_i$, s_i and $\widehat{\mu_3(\theta_i \mid \mathbf{y})}$ respectively we can then approximate the posterior distribution by matching moments with a Skew-Normal(ψ, ω, δ) distribution. A random variable, Y, having a Skew-Normal distribution with location, scale and skewness parameters, ψ, ω, and δ respectively, is defined in Azzalini (2005) and can be represented as

$$Y = \psi + \omega\left(\delta|Z_1| + \sqrt{(1-\delta^2)}Z_2\right)$$

where Z_1 and Z_2 are i.i.d. $N(0,1)$ and $-\infty < \psi < \infty$, $\omega > 0$ and $-1 \leq \delta \leq 1$. A Skew-Normal distribution is chosen over alternatives as it is characterized by three parameters and has support $(-\infty, \infty)$. A future extension of this work would be to also estimate the kurtosis and then moment match with a four parameter distribution such as the Skew-t.

We can now match the moments estimated for our random effects, θ_i, to those of the Skew-Normal by noting in Azzalini (2005) that

$$E(Y) = \psi + \omega\delta\sqrt{\frac{2}{\pi}} \tag{1.68}$$

$$\text{Var}(Y) = \omega^2(1 - \frac{2\delta^2}{\pi}) \tag{1.69}$$

$$\text{Skewness}(Y) = \frac{4-\pi}{2}\frac{\delta^3}{(\pi/2 - \delta^2)^{3/2}} \tag{1.70}$$

which leads to the following parameter estimates

$$\hat{\delta}_i \equiv \text{sign}(\gamma_i)\sqrt{\frac{\frac{\pi}{2}|\gamma_i|^{2/3}}{|\gamma_i|^{2/3}+((4-\pi)/2)^{2/3}}} \tag{1.71}$$

$$\hat{\omega}_i \equiv \sqrt{\frac{s_i}{(1-\frac{2\delta_i^2}{\pi})}} \tag{1.72}$$

$$\hat{\psi}_i \equiv \theta - \omega\delta_i\sqrt{\frac{2}{\pi}} \tag{1.73}$$

where we define the approximate skewness to be

$$\hat{\gamma}_i \equiv \frac{\widehat{\mu_3(\theta_i \mid \mathbf{y})}}{s_i^{3/2}}. \tag{1.74}$$

It is noted that the maximum magnitude of skewness for the Skew-Normal occurs when $\delta = \pm 1$ when all the weighting is on the skewed $|Z_1|$ and magnitude is thus equal to

$$\frac{4-\pi}{2}\frac{1}{(\pi/2-1)^{3/2}} \approx 0.9952. \tag{1.75}$$

Depending on the particular dataset it may occur that $|\hat{\gamma}_i|$ is greater than 0.9952. If this occurs we can redefine our estimated parameter estimate to be

$$\hat{\gamma}_i^* \equiv \text{sign}(\hat{\gamma}_i) \times \min\left(0.9952, |\hat{\gamma}_i|\right). \tag{1.76}$$

This allows for the maximum magnitude of skewness allowable by the Skew-Normal whilst still matching the mean and variance exactly and results in the matching distribution being a scale and location shift of a χ_1 random variable.

To assess the benefit of applying the Skew-Normal distribution we can undertake a simple simulation study and calculate the coverages under the Skew-Normal and Normal approximations. In Figure 1.8 and Figure 1.9 we present the results from a simulation study assuming $V_i = 1$ for all i, $\beta = 0$ and $k = 8, 20$. Data was generated according to these parameter values and for a sequence of values for the shrinkage factor, B. Intervals were then constructed for θ_i using both the Normal and the Skew-Normal approximations with mean, variance and skewness values noted in Sections 1.6.2 and 1.6.3 were used to construct the intervals.

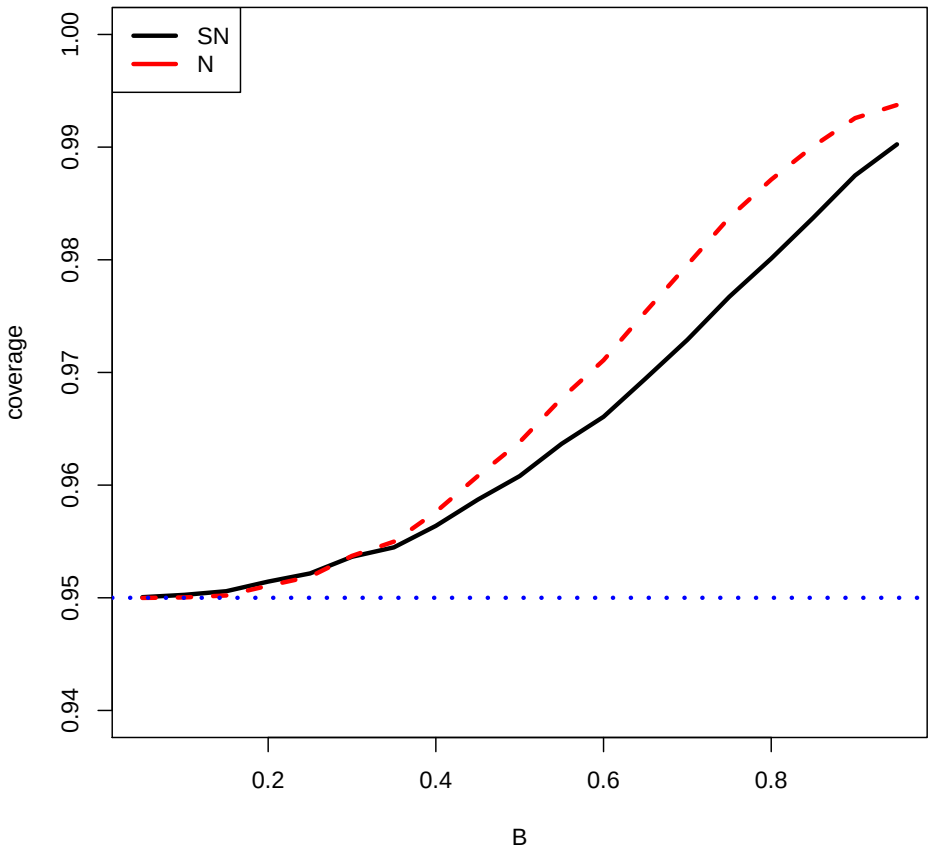

Figure 1.8: Coverages for the skewness simulation where k=8

From Figures 1.8 and 1.9 we see that the Skew-Normal approximation covers the

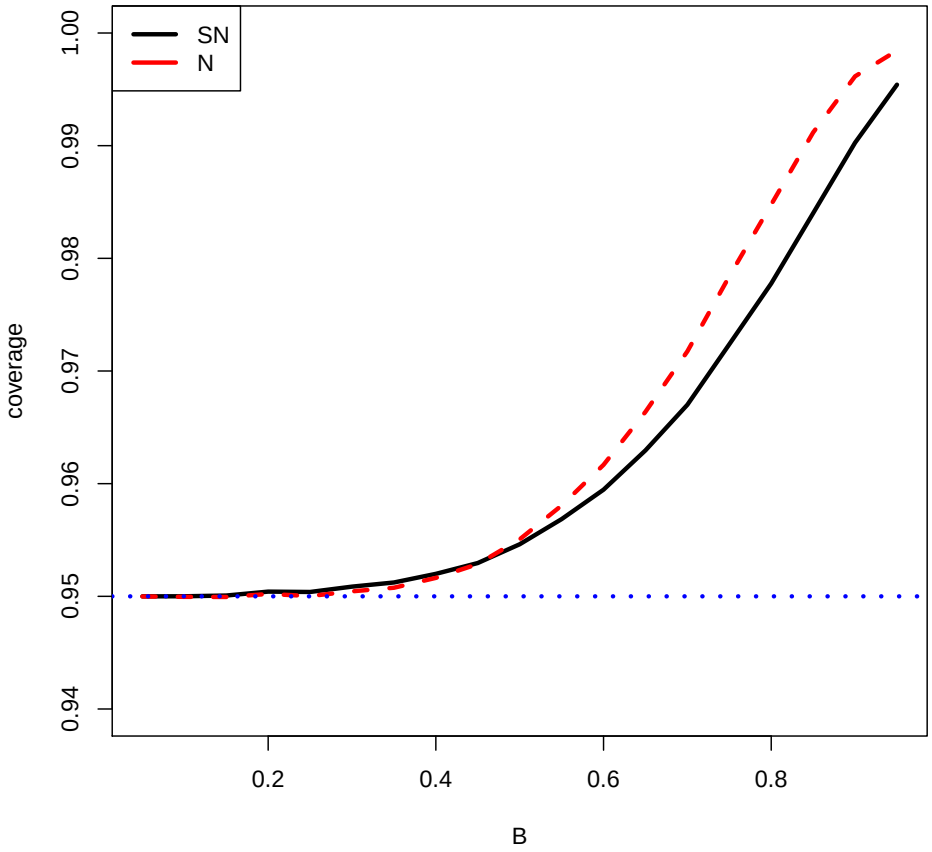

Figure 1.9: Coverages for the skewness simulation where k=20

required 95% and provides coverages closer to the reported nominal rate than the Normal approximation does.

1.7 Comparison with MCMC

In Section 1.6 we presented the methodology of how to estimate the Normal-Normal hierarchical model for the unequal variance case. In this section we will use this methodology to fit the three examples we have presented thus far and compare the results with those by fitting a Monte Carlo Markov Chain (MCMC). As we are primarily interested in estimating and constructing confidence intervals for the random effects, θ_i, the comparison demonstrates that our method produces results very similar to MCMC but due to it being a derivative based method can do so at a tiny fraction of the computational expense (see Section 2.3.5 and 2.4.3). The trade-off that we make for computational efficiency is that in GRIMM we are approximating the posterior by matching three moments to a Skew-Normal and so if quantities other than the point and interval estimates of θ_i are of interest then MCMC will be more appropriate.

1.7.1 Example: Schools

We revisit the schools data first presented in Section 1.3 and fit the model applying the methodology explained in Section 1.6. Figure 1.10 and Table 1.7 contain the results of the estimation.

From Figure Table 1.7 we note that unlike MLE and REML based methods which were shown in Section 1.3 to produce estimates of 100% shrinkage the GRIMM procedure produces shrinkage factors ranging from 0.41 to 0.73. This is corresponds to

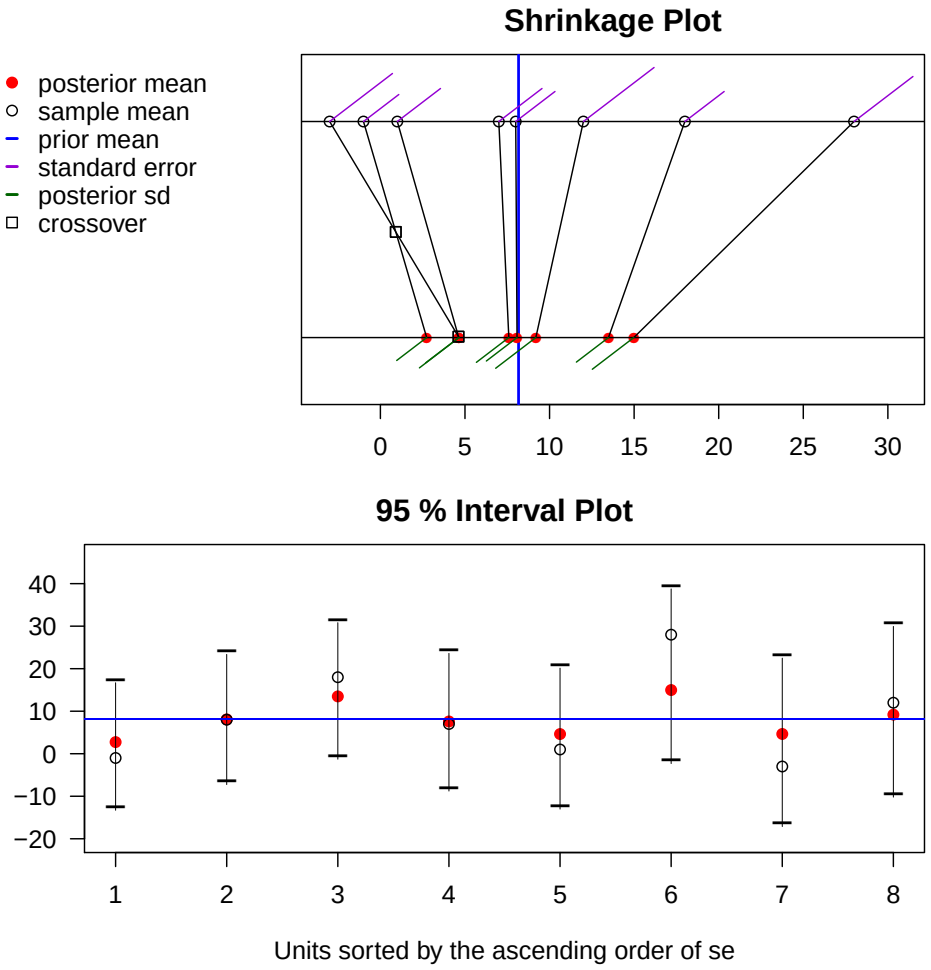

Figure 1.10: Plot of the results of the analysis of the schools data using GRIMM

an estimate of A of 117.92. To examine how well GRIMM fits this data Table 1.7 presents the results by using the MCMC algorithm described in Section 2.4.2.

Notice the remarkable similarities between MCMC(M) and GRIMM(G) in Table 1.7 suggesting that GRIMM is indeed a very good approximate procedure for MCMC. This is further demonstrated in Figure 1.11 where we compare the Skew-Normal approximation used by GRIMM with that from the samples obtained via the MCMC. In Figure 1.12 a closer look is taken at school 6 which we note has a significant amount

Table 1.7: Results of the analysis of the schools data with MCMC(M) and GRIMM(G)

i	y_i	$\sqrt{V_i}$	$\hat{B}_i^{(M)}$	$\hat{B}_i^{(G)}$	$\hat{\theta}_i^{(M)}$	$\hat{\theta}_i^{(G)}$	$s_i^{(M)}$	$s_i^{(G)}$	$\hat{\gamma}_i^{(M)}$	$\hat{\gamma}_i^{(G)}$
1	-1.00	9.00	0.48	0.41	3.33	2.71	7.34	7.63	-0.26	-0.28
2	8.00	10.00	0.52	0.46	8.00	8.06	7.49	7.82	0.02	-0.01
3	18.00	10.00	0.52	0.46	12.78	13.47	7.92	8.16	0.32	0.32
4	7.00	11.00	0.56	0.51	7.57	7.59	7.93	8.27	-0.03	-0.05
5	1.00	11.00	0.56	0.51	4.90	4.64	8.09	8.46	-0.23	-0.26
6	28.00	15.00	0.67	0.66	14.67	14.96	10.46	10.56	0.61	0.62
7	-3.00	16.00	0.69	0.69	4.62	4.68	9.98	10.10	-0.38	-0.41
8	12.00	18.00	0.73	0.73	9.13	9.17	10.26	10.22	0.17	0.14

of skewness which GRIMM is able to capture.

In Figures 1.11 and 1.12 the skewness present in the distributions of the random effects is apparent. The approximating Skew-Normal distribution seems to perform very well in capturing this skewness and indeed seems much more appropriate for this data than a Normal approximation. The 95% confidence intervals constructed using the three methods in Figure 1.12 are (-5.72, 35.68), (-2.32, 38.80) and (-2.91, 38.50) for the Normal, Skew-Normal and MCMC methods respectively. By incorporating the skewness present in the distribution the Skew-Normal approximation was better able to approximate the true distribution of random effects as shown by the similarities between the confidence intervals for the Skew-Normal and MCMC methods.

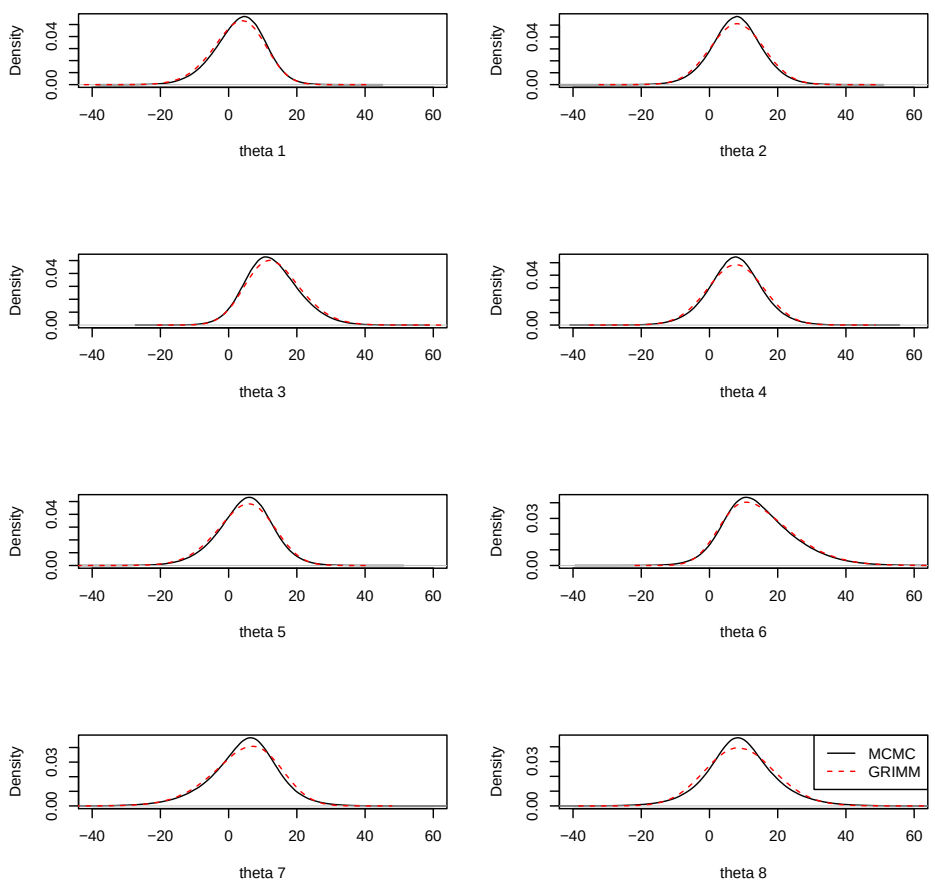

Figure 1.11: Plot of the comparison between GRIMM and MCMC for the eight schools data

1.7.2 Example: Hospitals

In Table 1.4 we presented the equal variance hospital data and fitted the model via direct posterior sampling noted in Table 1.5. For equal variances exact point estimates can be made as noted in (1.41) and (1.42), however, the data can still be analyzed using the methods for unequal variances and it gives us the opportunity to evaluate how well the GRIMM procedure fairs. In Table 1.8 and Figure 1.13 the results are presented from using the GRIMM procedure.

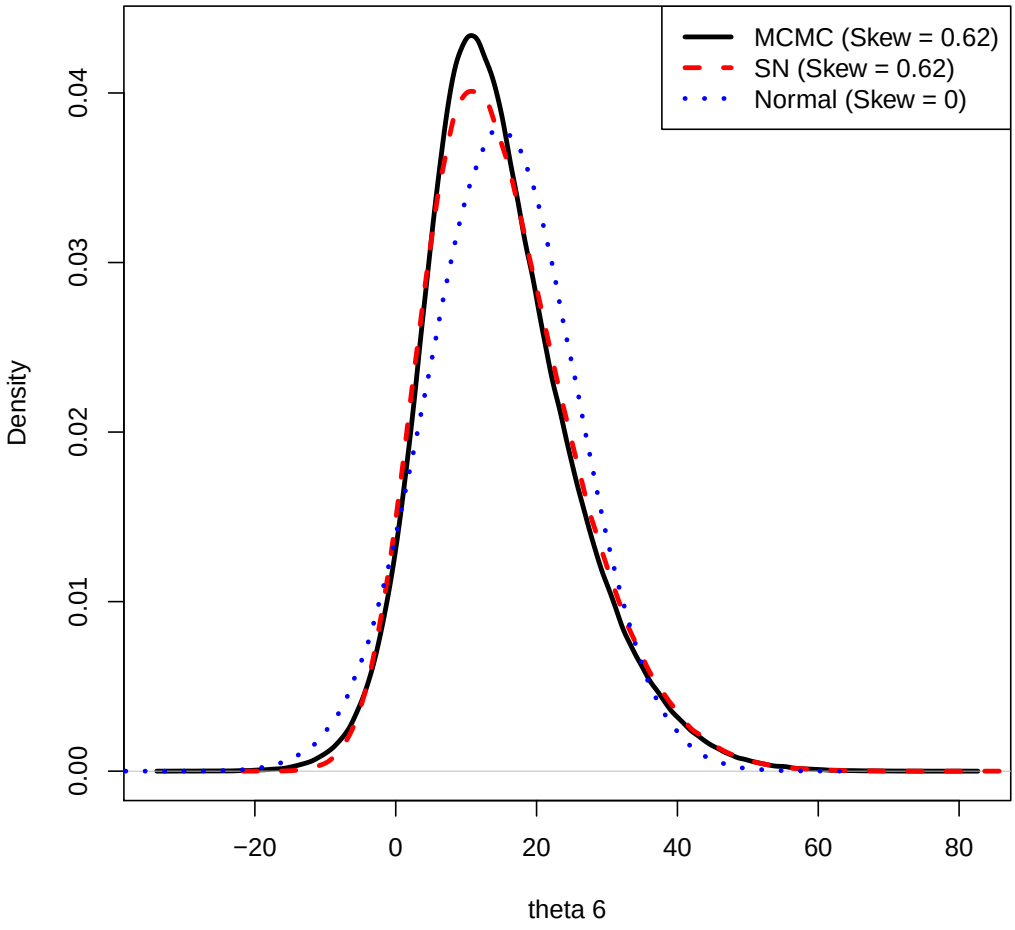

Figure 1.12: Plot of the comparison between GRIMM and MCMC for school 6

By examining Table 1.8 we note the similarities between the estimates given by GRIMM and those by direct posterior sampling. It is noted that GRIMM slightly underestimates the shrinkage factor, 0.36 for GRIMM compared with 0.42 via direct posterior sampling. This property of ADM being slightly conservative in nature was noted in Morris and Tang (2011). However, it is this conservative nature that keeps this derivative based method away from estimating shrinkages of 100% and as such is a justified trade-off.

As we did for the schools data, Figure 1.14 compares how the Skew-Normal ap-

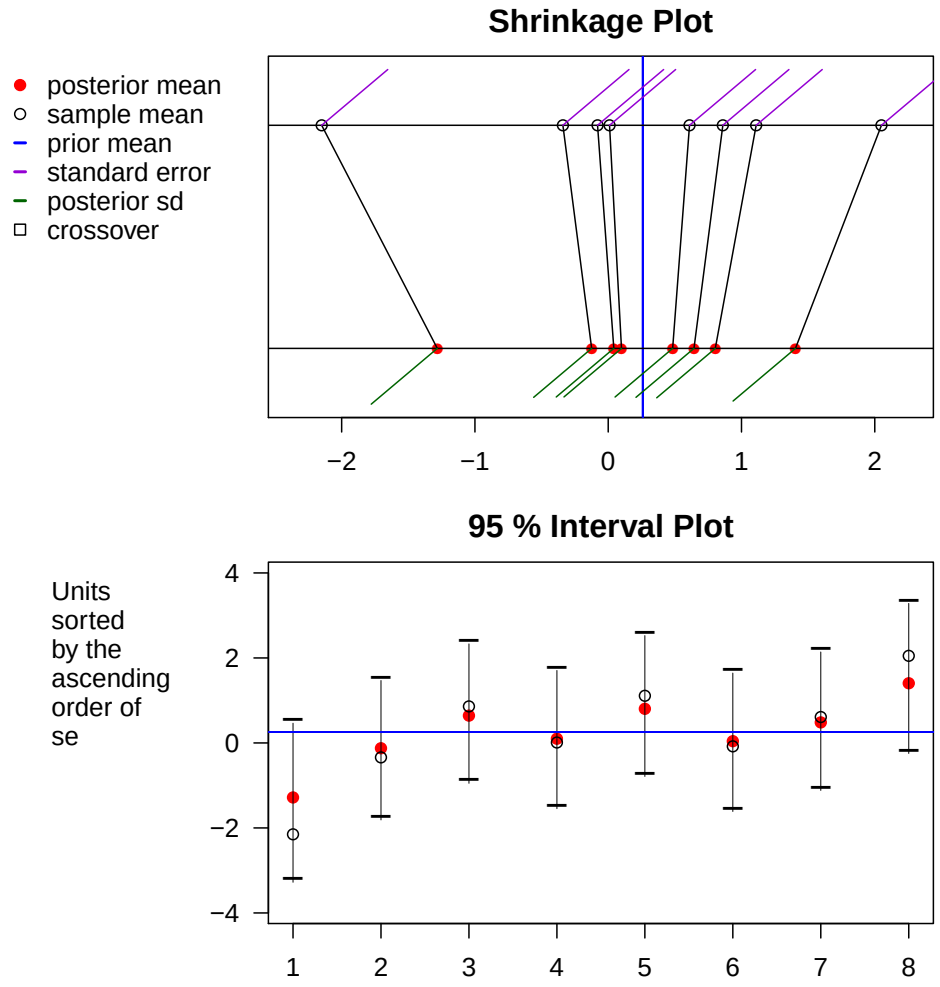

Figure 1.13: Results of the analysis of the equal variances hospital data with GRIMM

proximation given by GRIMM procedure compares with samples obtained via direct posterior sampling. Note that once again the GRIMM procedure appears to adequately approximate MCMC thus supporting evidence that GRIMM can be a reasonable alternative to MCMC.

GRIMM and MCMC comparisons were also made for the hospital data when variances were unequal. Due to brevity and the fact that the hospital data contains a relatively large number of groups (23) these results are presented in Appendix B. It can

Table 1.8: Results of the analysis of the equal variances hospital data with MCMC(M) and GRIMM(G)

i	y_i	$\sqrt{V_i}$	$\hat{B}_i^{(M)}$	$\hat{B}_i^{(G)}$	$\hat{\theta}_i^{(M)}$	$\hat{\theta}_i^{(G)}$	$s_i^{(M)}$	$s_i^{(G)}$	$\hat{\gamma}_i^{(M)}$	$\hat{\gamma}_i^{(G)}$
1	-2.15	1.00	0.42	0.36	-1.14	-1.28	0.96	0.95	-0.26	-0.27
2	-0.34	1.00	0.42	0.36	-0.09	-0.12	0.81	0.84	-0.15	-0.12
3	0.86	1.00	0.42	0.36	0.61	0.64	0.81	0.84	0.15	0.11
4	0.01	1.00	0.42	0.36	0.12	0.10	0.80	0.83	-0.07	-0.05
5	1.11	1.00	0.42	0.36	0.75	0.80	0.82	0.84	0.20	0.16
6	-0.08	1.00	0.42	0.36	0.06	0.04	0.80	0.83	-0.09	-0.06
7	0.61	1.00	0.42	0.36	0.46	0.48	0.80	0.83	0.09	0.07
8	2.05	1.00	0.42	0.36	1.30	1.40	0.89	0.90	0.29	0.27

be seen, however, that once again GRIMM proved to be a very good approximation to MCMC.

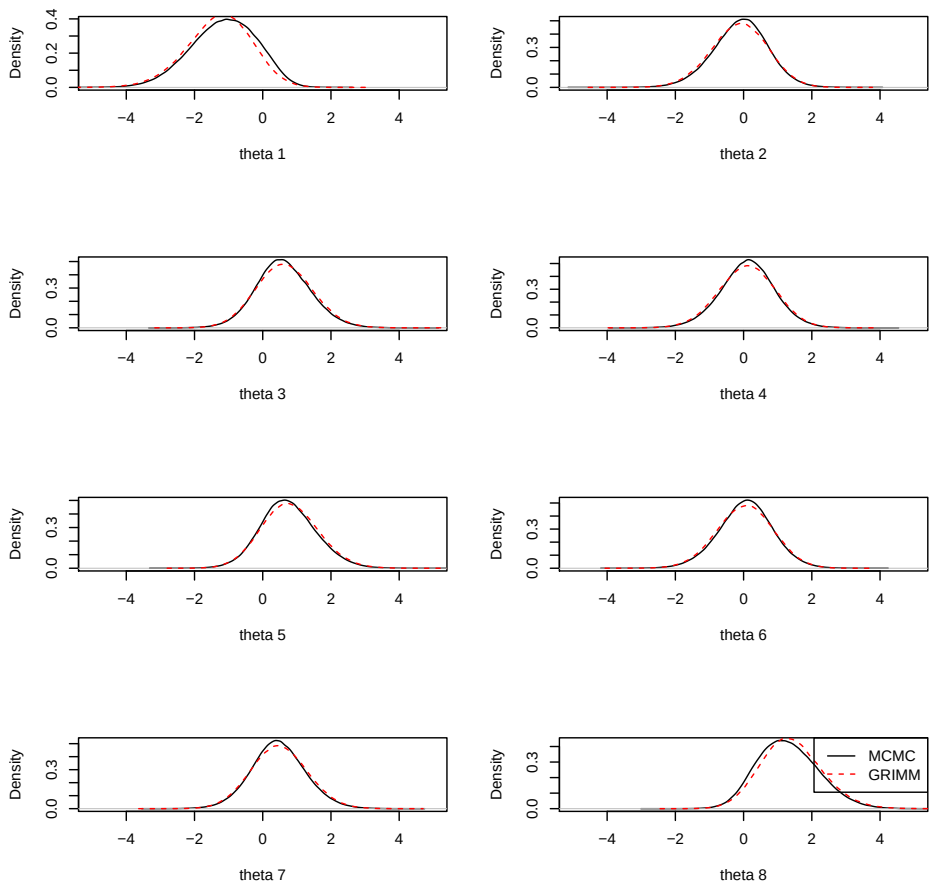

Figure 1.14: Plot of the comparison between GRIMM and direct posterior sampling for the equal variances hospital data

1.8 Coverage Evaluation

In Section 1.7 we examined how the GRIMM procedure fared compared to MCMC for the schools data and for the equal and unequal variance hospital data. Although reassuring that GRIMM produces estimates very similar to MCMC for these particular datasets it is also of interest to see how the procedure fairs over many similar datasets in terms of the frequency properties for the intervals of the random effects. In this section we will compare the frequency properties of GRIMM with those of MLE and REML by using a procedure described in Kelly et al. (2014b) as a method check.

The method check is a simulation based procedure applied after a model is fitted with GRIMM. Namely after A and β have been estimated by $\hat{A}$ and $\hat{\beta}$ respectively we can generate new data for each iteration j in the simulation such that

$$\theta_i^{(j)} \overset{ind}{\sim} N(x_i'\hat{\beta}, \hat{A}) \text{ for } i = 1 \dots k \tag{1.77}$$

$$y_i^{(j)} \mid \theta_i^{(j)} \overset{ind}{\sim} N(\theta_i^{(j)}, V_i) \text{ for } i = 1 \dots k. \tag{1.78}$$

Then for each iteration, j, we can again use GRIMM with the simulated data and construct $(1-\alpha)$ intervals, $(\hat{\theta}^{(j)}_{i,\alpha/2}, \hat{\theta}^{(j)}_{i,1-\alpha/2})$, and then calculate the probability the

interval covers the true value by calculating for each iteration j and unit i

$$\begin{aligned} p_i^{(j)} &\equiv P(\hat{\theta}_{i,\alpha/2}^{(j)} \leq \theta_i^{(j)} \leq \hat{\theta}_{i,1-\alpha/2}^{(j)}) \\ &= \Phi\left(\frac{\hat{\theta}_{i,1-\alpha/2}^{(j)} - ((1-\hat{B}_i)y_i^{(j)} + \hat{B}_i x_i'\hat{\beta})}{\sqrt{V_i(1-\hat{B}_i)}}\right) \\ &- \Phi\left(\frac{\hat{\theta}_{i,\alpha/2}^{(j)} - ((1-\hat{B}_i)y_i^{(j)} + \hat{B}_i x_i'\hat{\beta})}{\sqrt{V_i(1-\hat{B}_i)}}\right). \end{aligned} \tag{1.79}$$

Averaging $p_i^{(j)}$ over many simulations thus allows us to evaluate the frequency properties of the GRIMM procedure for a particular dataset. It is important to point out that the method check is designed to check that the method of inference provides the stated frequency properties at particular values of the hyperparameters, (β, A), assuming the model is correct. In the context of this paper it is checking that the formal Bayes interval coverages (e.g. 95%) do not exceed the frequency confidence. It serves a different purpose than model checking and is to be used in addition to model checking procedures such as posterior predictive checks. As a method check requires the model to be estimated multiple times the speed of the estimation procedure is extremely important. The speed of GRIMM allows a method check to be conducted for datasets like those presented here within 1 second whereas running a MCMC at each iteration of the method check is usually not feasible given reasonable computational constraints.

In Figure 1.15 we present the method check results for the three example datasets noted in Section 1.7 for 95% intervals. It is noted that whilst MLE and REML based

procedures undercover in every example the GRIMM procedure provides at least 95% coverage for each unit for each dataset. This fact strongly supports the idea that GRIMM is an ideal procedure if good frequency properties are desired and should be used over other MLE and REML based procedures.

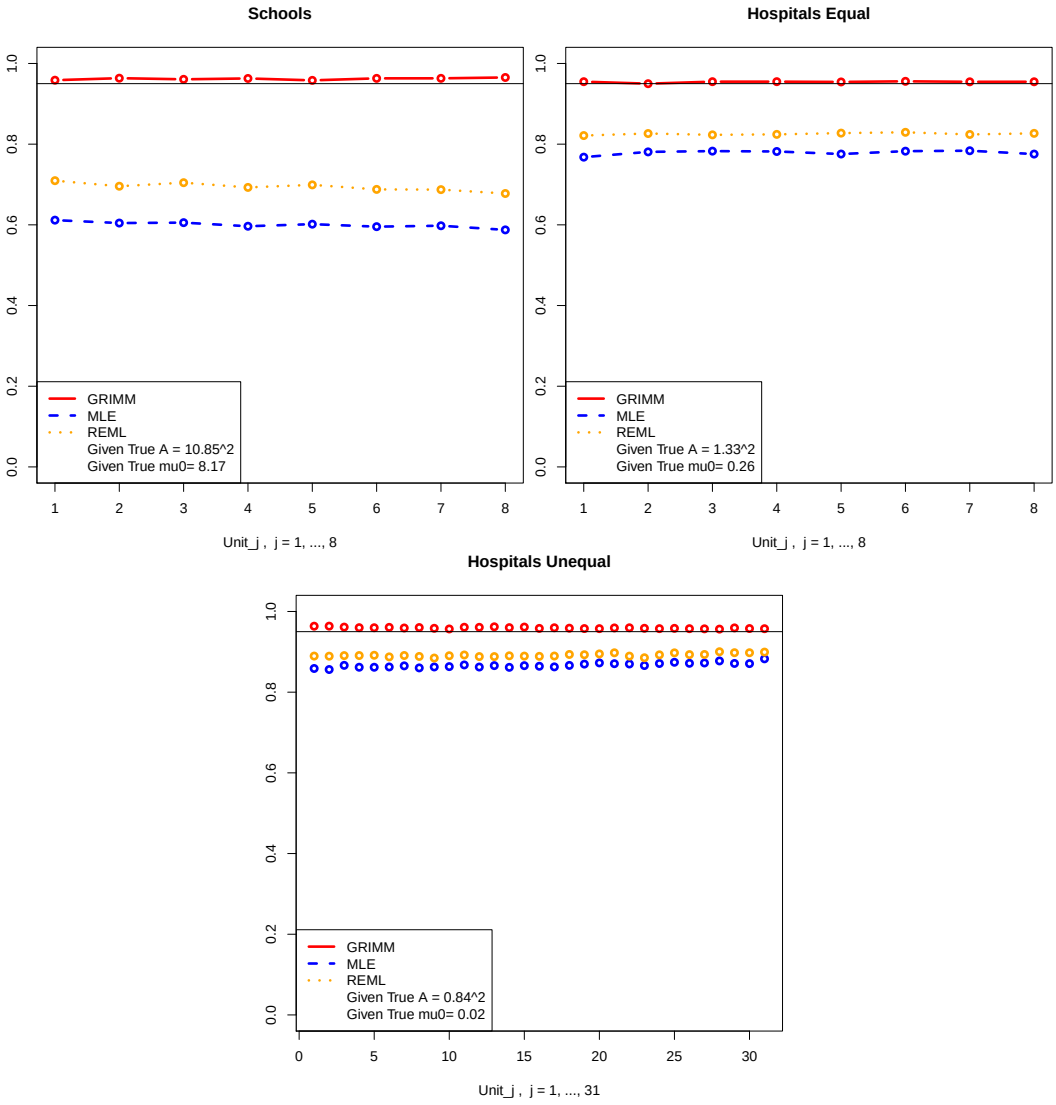

Figure 1.15: Coverage comparisons between GRIMM, MLE and REML for the three example datasets

1.9 Rgbp

Rgbp (Kelly et al. (2014a)) is a free software package for the R (R Core Team (2014)) programming language available on CRAN (`http://cran.r-project.org/web/packages/Rgbp/index.html`). Rgbp, an acronym for R-Gaussian-Binomial-Poisson, is a package which not only fits the Normal-Normal hierarchical model (GRIMM), noted in this dissertation, but also the Binomial-Beta (BRIMM) and Poisson-Gamma (PRIMM) hierarchical models. The software was used throughout this dissertation to produce the results in Tables 1.7, 1.8 and B.1 and Figures 1.10, 1.8 and B.1. The package can also provide the method check procedure referenced in Section 1.8 albeit without the MLE and REML comparisons. For more information about the package the reader is guided to Kelly et al. (2014a) and Kelly et al. (2014b)

1.10 Discussion and Conclusions

In this chapter we have presented the Normal-Normal hierarchical model and proposed a method of estimation known as Gaussian Regression Interactive Multilevel Modeling (GRIMM). Although not presented in the main text if the reader is interested in an historical perspective of the Normal-Normal hierarchical model they are referred to Appendix A.

In Section 1.2 we presented the inferential and descriptive versions of the model and then continued in Section 1.4 to assume priors which we later show in Section 1.8 led to good frequency properties. We proved the posterior propriety for our assumed priors and also proved when posterior propriety was feasible for a class of priors for the unequal variance case with covariates if an alternative prior is desired.

In Section 1.6 we presented the procedure known as Adjustment for Density Maximization which allows for estimating the model via differentiation and aims to rectify issues based on maximum likelihood estimation that are shown to distort shrinkages and random effect inferences and showed via the schools example why MLE or REML estimation may lead to 100% shrinkages and poor interval estimation. We then extended the work of Morris and Tang (2011) to incorporate skewness in the approximation of the random effects which we demonstrated to be a desirable extension.

In Section 1.7 we compared GRIMM with inferences based on MCMC and showed for the examples presented that GRIMM produces comparable results. A benefit of

GRIMM over MCMC is that it is a method that can be estimated using derivatives via Newton-Raphson or via EM based algorithms that are presented in Chapter 2. Another benefit of GRIMM is that it offers a sense of reproducibility. Every example presented here can be reproduced using the R package Rgbp Kelly et al. (2014a) and the results will be exactly the same. MCMC on the other hand suffers from Monte Carlo errors and thus without using the same programming language and the same seed to implement the MCMC the results may differ from user to user. Additionally GRIMM is more stable in the estimation procedure in that it is often easier to check if convergence of a Newton-Raphson or EM based algorithm has converged than if a MCMC has had a large enough burn-in and began to produce draws from the stationary distribution.

Due to the difficulty in checking convergence of a MCMC and given that typically, derivative and EM based algorithms converge must faster than any MCMC procedure it can be said that GRIMM would be an ideal candidate to be used for simulation purposes where a Normal-Normal hierarchical model would needed to be fit numerous times. This is evident in Section 1.8 where we introduced the concept of a method check where the speed of GRIMM allows a method check to be conducted for most datasets within 1 second whereas running a MCMC at each iteration of the method check is usually not feasible given reasonable computational restraints.

Of course, MCMC, definitely has its advantages and in Chapter 2 we present a data-augmentation based Gibbs sampler to fit the Normal-Normal hierarchical model. A benefit of MCMC is that as long as the Markov chain has converged its approxi-

mation error is only as large as the Monte Carlo error and can be reduced by drawing further samples. In the examples presented we showed that the Skew-Normal distribution is an extremely good approximation to the posterior distribution of the random effects, however, this was only accomplished by comparing with the posterior obtained via MCMC. There may exists other quantities of interest regarding the random effects (such as extreme quantile values) where a distribution that only matches on three moments may not be an appropriate approximation and it may be difficult to determine how good the approximation is without implementing an MCMC. If that is case MCMC may very well be a better option.

GRIMM can also be used in conjunction with MCMC. For example, the estimates from GRIMM can be used to initialize the MCMC with the hope that it will converge faster. It can also be used as a first pass through to fit the model to get an approximate idea of what the posterior distribution looks like for the random effects and depending on the users purposes if it is worth implementing an MCMC or not.

In Section 1.8 we compared the GRIMM's frequency properties with MLE and REML based inferences. It was noted that MLE and REML often have coverage well below the desired and reported value. This is largely due to the fact that MLE and REML can lead to 100% shrinkages and that the Normal approximation is often not appropriate for small samples. GRIMM rectifies this by using ADM in the estimation procedure and allows for skewness in the posterior of the random effects by approximating the distribution with a Skew-Normal. An avenue of future research would be to extend GRIMM to also estimate the Kurtosis and match a four parameter

distribution such as the NEF-CHS or Skew-t distribution (Morris and Lock (2009)).

It is worth noting that GRIMM is designed to improve on MLE and REML based inferences. Unlike MCMC, which provides draws from the posterior distribution, GRIMM focuses on providing point and interval estimates for the random effects, θ_i, but does so at a fraction of the computation expense (often 1000 times faster than an MCMC with similar approximation error). This speed allows for method checking which may or may not be computationally feasible under MCMC and so the choice of GRIMM over MCMC really depends on what quantities are of interest and if a frequency evaluation is required.

In conclusion, GRIMM is a procedure that utilizes Bayesian machinery to produce inferences with desirable frequency properties. The procedure is readily available via the R package Rgbp (Kelly et al. (2014a)) and is a preferred method of estimation to MLE and REML based methods and may be used as a very reasonable alternative to MCMC.

Chapter 2

Data Augmentation Estimation Methods

Preface

This entirety of this chapter is under the supervision and in collaboration with my advisor Professor Carl Morris.

2.1 Introduction

In Chapter 1 we presented the Normal-Normal hierarchical model and suggested a framework for modeling and estimation known as GRIMM. In this chapter we present an algorithm for estimating GRIMM known as Adjustment for Density Maximiza-

tion Data Augmentation (ADMDA) and explore its theoretical and computational properties.

It was noted in Section 1.5 that when equal variances are present the inference for the hyperparameters, β and A, becomes much more tractable due to the symmetry present in the data. Unfortunately, this symmetry is lost when moving to the unequal variance case and thus in this chapter we present a data augmentation method similar to that in Van Dyk and Tang (2003) in order to utilize the symmetry present in the complete data to help estimate the unequal variance problem. The data augmentation methodology is explained in Section 2.2 and useful distributions related to the data augmentation procedure are derived.

In Section 2.3 we present the EM-based algorithms to fit the Normal-Normal hierarchical for MLE and REML based estimation for both the data augmentation scheme and the scheme where the θ_i's are treated as missing which we have dubbed the traditional EM method. As shown throughout Chapter 1, MLE and REML based methods provide interval estimates with poor frequency properties and as such we propose a new algorithm for estimation combining the ADM procedure with the data augmentation method which we name Adjustment for Density Maximization Data Augmentation (ADMDA).

In Section 2.3.4 some of the properties of the algorithms under the data augmentation scheme are derived and in particular it is shown that the ADMDA algorithm does not decrease the adjusted likelihood at each iteration of the algorithm.

In Section 2.3.5 we conduct two simulation studies and tackle a real data example

to make a few performance comparisons between ADMDA and the other algorithms presented in Section 2.3.

Finally it is noted in Section 2.4 that the data augmentation scheme also has benefits if undertaking MCMC. We present a Gibbs algorithm based on this data augmentation scheme and compare with an alternative Gibbs algorithm based on their performance for fitting the example datasets present in Section 1.3 and Section 1.5.3.

2.2 Data Augmentation Methodology

As in Chapter 1 the model for the observed data with unequal variances is

$$y_i^{obs} \mid \theta_i \overset{ind}{\sim} N(\theta_i, V_i) \text{ for } i = 1 \ldots k$$

$$\theta_i \mid \phi \overset{ind}{\sim} N(x_i^{'}\beta, A) \text{ for } i = 1 \ldots k$$

where $\phi \equiv (\beta, A)$, $V_1 \leq V_2 \leq \ldots \leq V_k$, X is the $k \times r$ matrix of covariates, $x_i^{'}$ is the i^{th} row of X and $\Theta = (\theta_1, \ldots, \theta_k)$.

It is noted that the lack of symmetry between groups, i, is due to the unequal variances. To try and rectify this we define for $i = 1 \ldots k$,

$$\begin{aligned} w_i &= 1 - V_1/V_i & (2.1)\\ y_i^{mis}|\theta_i &\sim N(\theta_i, \frac{V_1}{w_i}) & (2.2)\\ y_i^{com} &= (1 - w_i)y_i^{obs} + w_i y_i^{mis}. & (2.3) \end{aligned}$$

Then by construction

$$y_i^{com} \mid \theta_i \sim N(\theta_i, V_1) \tag{2.4}$$

and

$$y_i^{com} \mid \phi \overset{ind}{\sim} N(x_i^{'}\beta, V_1 + A). \tag{2.5}$$

Intuitively this can method can be understood by imagining that the sample mean for each group i, is based off a certain number of individuals within that group and

that the variances can be thought of $V_i = \sigma^2/n_i$. Due to the fact each has a different number of members, $n_1 \geq n_2 \ldots \geq n_k$, the resulting variances are different. We then construct the hypothetical scenario where we imagine each group had the same number of members, n_1, and that a certain fraction, w_i, of those members observations are missing. It can be seen from (2.4) that the complete data has equal variances and we can exploit this symmetry to make inferences more tractable.

In order to make any inferences about, θ_i, or our hyperparameters (β, A) we must first derive some useful conditional distributions involving the complete and observed data, $\mathbf{y}^{com}$ and $\mathbf{y}^{obs}$ respectively.

From (1.33) we have

$$\beta \mid \hat{\beta}^{com}, A \sim N(\hat{\beta}^{com}, (V_1 + A)(X'X)^{-1}) \tag{2.6}$$

where $\hat{\beta}^{com} \equiv (X'X)^{-1}X'\mathbf{y}^{com}$

and from (1.34) we have

$$p(A \mid S^{com}) \propto (V_1 + A)^{-\frac{k-r}{2}} e^{-\frac{S^{com}}{2(V_1+A)}} \tag{2.7}$$

where $S^{com} \equiv \sum_{i=1}^{k} (y_i^{com} - x_i'\hat{\beta}^{com})^2$

and from (1.35) we have

$$\theta_i \mid y_i^{com}\beta, A \sim N((1 - B_k)y_i^{com} + B_k x_i'\beta, V_1(1 - B_k)). \tag{2.8}$$

Perhaps more interesting is the distribution of $y_i^{com} \mid y_i^{obs}, \beta, A$ from (2.3) we see that y_i^{com} is a linear combination of Normal random variables and thus itself will be Normal. To find its mean and variance we can apply Adam's and EVE's law

$$\begin{aligned} E(y_i^{com} \mid y_i^{obs}, \beta, A) &= E(E(y_i^{com} \mid \theta_i, y_i^{obs}\beta, A)|y_i^{obs}, \beta, A) \\ &= E(E((1-w_i)y_i^{obs} + w_i y_i^{mis} \mid \theta_i, y_i^{obs}, \beta, A)|y_i^{obs}, \beta, A) \\ &= E((1-w_i)y_i^{obs} + w_i\theta_i \mid y_i^{obs}\beta, A) \\ &= (1-w_iB_i)y_i^{obs} + w_iB_ix_i^{'}\beta \end{aligned} \tag{2.9}$$

$$\begin{aligned} \mathrm{Var}(y_i^{com} \mid y_i^{obs}, \beta, A) &= E(\mathrm{Var}(y_i^{com} \mid \theta_i, y_i^{obs}\beta, A)|y_i^{obs}, \beta, A) \\ &\quad + \mathrm{Var}(E(y_i^{com} \mid \theta_i, y_i^{obs}, \beta, A) \mid y_i^{obs}, \beta, A) \\ &= E(\mathrm{Var}((1-w_i)y_i^{obs} + w_i y_i^{mis} \mid \theta_i, y_i^{obs}, \beta, A) \mid y_i^{obs}, \beta, A) \\ &\quad + \mathrm{Var}((1-w_i)y_i^{obs} + w_i\theta_i|y_i^{obs}, \beta, A) \\ &= w_iV_1 + w_i^2(V_i(1-B_i)) \end{aligned} \tag{2.10}$$

and therefore

$$y_i^{com} \mid y_i^{obs}, \beta, A \overset{ind}{\sim} N((1-w_iB_i)y_i^{obs} + w_iB_ix_i^{'}\beta, w_iV_1 + w_i^2(V_i(1-B_i))). \tag{2.11}$$

Although the full conditional in (2.11) is useful in many algorithms such as the Gibbs sampler for some of the EM algorithms proposed in Section 2.3 the distribution

of $y_i^{com} \mid y_i^{obs}, A$ is necessary. Utilizing Adam's and EVE's laws again we see that

$$\begin{aligned} E(y_i^{com} \mid y_i^{obs}, A) &= E(E(y_i^{com} \mid y_i^{obs}, A, \beta) \mid y_i^{obs}, A) \\ &= (1 - w_i B_i) y_i^{obs} + w_i B_i x_i^{'} \hat{\beta}_A \end{aligned} \tag{2.12}$$

$$\begin{aligned} \text{Var}(y_i^{com} \mid y_i^{obs}, A) &= E(\text{Var}(y_i^{com} \mid y_i^{obs}, A, \beta) \mid y_i^{obs}, A) \\ &\quad + \text{Var}(E(y_i^{com} \mid y_i^{obs}, A, \beta) | y_i^{obs}, A) \\ &= w_i V_1 + w_i^2 (V_i (1 - B_i)) + w_i^2 B_i^2 x_i^{'} \Sigma_A x_i. \end{aligned} \tag{2.13}$$

Due to the normality of $\beta | y_i^{obs}, A$ as evidence by (1.17) we have that

$$\begin{aligned} y_i^{com} \mid y_i^{obs}, A \sim N((1 - w_i B_i) y_i^{obs} + w_i B_i x_i^{'} \hat{\beta}_A, \\ w_i V_1 + w_i^2 (V_i (1 - B_i)) + w_i^2 B_i^2 x_i^{'} \Sigma_A x_i). \end{aligned} \tag{2.14}$$

In Appendix C we tentatively propose a method to obtain a posterior mean estimate of the shrinkage factor, B_1, based on this data augmentation scheme. This proposal, however, is an area of active research and as such we did not feel comfortable including it in the main text until we are confident of its properties and viability.

2.3 EM Based Algorithms

To deal with the difficulty in using the data augmentation scheme to find exact point estimates in this section we suggest some EM (Dempster et al. (1977a)) based algorithms based on the data augmentation scheme to achieve MLE, REML and ADM estimates. We also present the traditional EM procedures for this type of hierarchical model and compare the methods based on convergence and computation efficiency.

2.3.1 Traditional EM

The traditional EM method of estimation for the Normal-Normal model does not rely on the data augmentation in Section 2.2 but instead treats the θ_i's as the missing data and does not integrate them out. Here we present the EM algorithms for fitting the Normal-Normal hierarchical model via MLE and REML and will later compare these to algorithms based on the data augmentation scheme.

Maximum Likelihood Estimation

Defining $\phi \equiv (\beta, A)$ the complete data log-likelihood up to a constant is

$$ll_{MLE}(\phi) \equiv -k/2\log(A) - \frac{\sum_{i=1}^{k}(\theta_i - x_i'\beta)^2}{2A}. \tag{2.15}$$

E-step

$$Q_{MLE}(\phi \mid \phi^{(t)}) \equiv E(\log(L(\phi)) \mid \mathbf{y}\phi^{(t)}) \tag{2.16}$$

$$= -k/2\log(A) - \frac{\sum_{i=1}^{k} E((\theta_i - x_i'\beta)^2|\mathbf{y}, \phi^{(t)})}{2A} \tag{2.17}$$

where the expectation is over the distribution of $\theta_i \mid \mathbf{y}, \phi^{(t)}$ and

$$E((\theta_i - x_i'\beta)^2 \mid \mathbf{y}, \phi^{(t)}) = (E(\theta_i|\mathbf{y}, \phi^{(t)}) - x_i'\beta)^2 + \mathrm{Var}(\theta_i \mid \mathbf{y}, \phi^{(t)}) \tag{2.18}$$

where

$$E(\theta_i \mid \mathbf{y}, \phi^{(t)}) = (1 - B_i^{(t)})y_i + B_i^{(t)}x_i'\beta^{(t)} \tag{2.19}$$

$$\mathrm{Var}(\theta_i \mid \mathbf{y}, \phi^{(t)}) = V_i(1 - B_i^{(t)}) \tag{2.20}$$

M-step By maximizing $Q(\phi \mid \phi^t)$ we see that

$$\beta^{(t+1)} \equiv (X'X)^{-1}X'E(\Theta|\mathbf{y}, \phi^{(t)}) \tag{2.21}$$

$$A^{(t+1)} \equiv \frac{\sum_{i=1}^{k} E((\theta_i - x_i'\beta^{(t+1)})^2 \mid \mathbf{y}, \phi^{(t)})}{k} \tag{2.22}$$

Restricted Maximum Likelihood Estimation

The restricted complete data log-likelihood up to a constant is

$$ll(A)_{REML} \equiv -\frac{k-r}{2}\log(A) - \frac{S_\theta}{2A} \tag{2.23}$$

where $S_\theta \equiv \sum_{i=1}^{k}(\theta_i - x_i'\hat{\beta}_\theta)^2$ and $\hat{\beta}_\theta \equiv (X'X)^{-1}X'\Theta$.

E-step

$$\begin{aligned} Q_{REML}(A \mid A^{(t)}) &\equiv E(ll(A)_{REML} \mid \mathbf{y}, A^{(t)}) \\ &= -\frac{k-r}{2}\log(A) - \frac{E(S_\theta \mid \mathbf{y}, A^{(t)})}{2A} \end{aligned} \tag{2.24}$$

$$\begin{aligned} E(S_\theta|\mathbf{y}^{obs}, A^{(t)}) &= E\left((\Theta - X\hat{\beta}_\theta)'(\Theta - X\hat{\beta}_\theta) \mid \mathbf{y}, A^{(t)}\right) \\ &= E\left((\Theta - H\Theta)(\Theta - H\Theta)' \mid \mathbf{y}, A^{(t)}\right) \\ &= E(\Theta'|\mathbf{y}, A^{(t)})(I-H)E(\Theta \mid \mathbf{y}, A^{(t)}) + \mathrm{tr}((I-H)D_\theta) \end{aligned} \tag{2.25}$$

where H is the hat matrix $H \equiv X(X'X)^{-1}X'$ and $D_\theta \equiv \mathrm{diag}(\mathrm{Var}(\theta_i \mid \mathbf{y}, A^{(t)}))$. Note that $E(\theta_i \mid \mathbf{y}, A^{(t)})$ and $\mathrm{Var}(\theta_i \mid y_i, A^{(t)})$ are given in (1.53) and (1.54) respectively.

M-step By maximizing $Q_{REML}(A \mid A^{(t)})$ we see that

$$A^{(t+1)} \equiv \frac{E(S_\theta \mid \mathbf{y}, A^{(t)})}{k-r}. \tag{2.26}$$

2.3.2 Data Augmentation EM Algorithms

Here we utilize the data augmentation scheme presented in Section 2.2 to implement the EM algorithm to undertake estimation of the Normal-Normal hierarchical model via maximum likelihood and REML. Additionally we propose and an EM-based algorithm we name Adjustment for Density Maximization Data Augmentation (ADMDA) that incorporates the ADM procedure with the data augmentation setup.

Even though ADM is the preferred method of estimation due to the reasons mentioned in Chapter 1 we present algorithms for MLE and REML in order to compare how the data augmentation scheme compares with traditional EM based methods noted in Section 2.3.1.

Maximum Likelihood Estimation

The Q-function is the expectation of the complete data log-likelihood where the expectation is with respect to the distribution of $\mathbf{y}^{com} \mid \mathbf{y}^{obs}, \phi = \phi^{(t)}$

$$\begin{aligned} Q_{MLE}(\phi \mid \phi^{(t)}) &\equiv \sum_{i=1}^{k} E(\log(f(y_i^{com} \mid \phi)) \mid y_i, \phi^{(t)}) \\ &= -\frac{k}{2}\log(V_1 + A) \\ &- \frac{\sum_{i=1}^{k} E((y_i^{com} - x_i^{'}\beta)^2 | y_i, \phi^{(t)})}{2(V_1 + A)} + C \end{aligned}$$

E-step

$$\begin{aligned} E(y_i^{com} \mid y_i^{obs}, \phi = \phi^{(t)}) &= (1 - w_i B_i^{(t)}) y_i^{obs} + w_i B_i^{(t)} x_i^{'} \beta^{(t)} \\ \mathrm{Var}(y_i^{com} \mid y_i^{obs}, \phi = \phi^{(t)}) &= w_i V_1 + w_i^2 (V_i (1 - B_i^{(t)}) \\ E((y_i^{com} - x_i^{'}\beta)^2 | y_i^{obs}, \phi = \phi^{(t)}) &= (E(y_i^{com} | y_i^{obs}, \phi = \phi^{(t)}) - x_i^{'}\beta)^2 \\ &+ \mathrm{Var}(y_i^{com} | y_i^{obs}, \phi = \phi^{(t)}). \end{aligned}$$

M-step

$$\hat{\beta}^{(t+1)} \equiv (X^{'}X)^{-1}X^{'}E(\mathbf{y}^{com} \mid \mathbf{y}^{obs}, \phi = \phi^{(t)}) \tag{2.27}$$

$$\hat{A}^{(t+1)} \equiv \max\left\{\frac{\sum_{i=1}^{k} E((y_i^{com} - x_i^{'}\beta^{(t+1)})^2 | y_i^{obs}, \phi = \phi^{(t)})}{k} - V_1, 0\right\} \tag{2.28}$$

Restricted Maximum Likelihood Estimation

From (2.7) we have that the complete data log-likelihood is

$$ll_{com}(A) \equiv -\frac{k-r}{2}\log(V_1 + A) - \frac{S^{com}}{2(V_1 + A)} \tag{2.29}$$

where $S^{com} \equiv \sum_{i=1}^{k}(y_i^{com} - x_i'\hat{\beta}^{com})^2$

The Q-function is the expectation of the complete data log-likelihood, $ll_{com}(A)$, where the expectation is with respect to the distribution of $\mathbf{y}^{com} \mid \mathbf{y}^{obs}, A^{(t)}$.

$$Q(A \mid A^{(t)}) \equiv -\frac{k-r}{2}\log(V_1 + A) - \frac{E(S^{com}|\mathbf{y}^{obs}, A^{(t)})}{2(V_1 + A)} \tag{2.30}$$

E-step

$$\begin{aligned} E(S^{com}|\mathbf{y}^{obs}, A^{(t)}) &= \sum_{i=1}^{k} E((y_i^{com} - x_i'\hat{\beta}^{com})^2 \mid \mathbf{y}^{obs}, A^{(t)}) \\ &= E\left((\mathbf{y}^{com} - X\hat{\beta}^{com})'(\mathbf{y}^{com} - X\hat{\beta}^{com}) \mid \mathbf{y}^{obs}, A^{(t)}\right) \\ &= E\left((\mathbf{y}^{com} - H\mathbf{y}^{com})(\mathbf{y}^{com} - H\mathbf{y}^{com})' \mid \mathbf{y}^{obs}, A^{(t)}\right) \\ &= E((\mathbf{y}^{com})' \mid \mathbf{y}^{obs}, A^{(t)})(I - H)E((\mathbf{y}^{com}) \mid \mathbf{y}^{obs}, A^{(t)}) \\ &+ \mathrm{tr}((I - H)D_{com}) \end{aligned} \tag{2.31}$$

where H is the hat matrix $H \equiv X(X'X)^{-1}X'$ and $D_{com} \equiv \mathrm{diag}(\mathrm{Var}(y_i^{com} \mid \mathbf{y}, A^{(t)}))$. Note that $E(y_i^{com} \mid y_i^{obs}, A^{(t)})$ and $\mathrm{Var}(y_i^{com} \mid y_i^{obs}, A^{(t)})$ are given in (2.12) and (2.13) respectively.

M-step Denoting $S_* \equiv E(S^{com}|\mathbf{y}^{obs}, A^{(t)})$ found in the E-step. We have that

$$\frac{\partial Q}{\partial A} = \frac{S_*}{2(A + V_1)^2} - \frac{k - r}{2(A + V_1)} \tag{2.32}$$

Solving $\frac{\partial Q}{\partial A} = 0$ we get

$$A^{(t+1)} \equiv \max\left\{\frac{S_*}{k - r} - V_1, 0\right\}. \tag{2.33}$$

2.3.3 Adjustment for Density Maximization Data Augmentation (ADMDA)

The EM algorithm in Section 2.3.2 can be altered so that instead of MLE or REML based inferences ADM can be used. In the Normal-Normal hierarchical model the random effects, θ_i, have mean and variance linear in the shrinkage factor, B_i. For our purposes the random effects are the quantity of interest and as such it desirable to estimate A such that the means and variance estimates of the random effects are unbiased. It is noted that B_i is a convex function of A and so, unlike MLE and REML, ADM is desirable as it aims to provide an estimate, $\hat{A}$, such that $\hat{B}_i = \frac{V_i}{V_i+\hat{A}} \approx E(B_i|\mathbf{y})$.

From (2.7) we have that the complete data adjusted log likelihood is

$$ll_{adj}(A) \equiv \log(A) - \frac{k-r}{2}\log(V_1 + A) - \frac{S^{com}}{2(V_1 + A)} \tag{2.34}$$

where $S^{com} \equiv \sum_{i=1}^{k} (y_i^{com} - x_i'\hat{\beta}^{com})^2$

The adjusted Q-function is the expectation of the complete data adjusted log-likelihood, $ll_{adj}(A)$, where the expectation is with respect to the distribution of $\mathbf{y}^{com} \mid \mathbf{y}^{obs}, A^{(t)}$.

$$Q_{adj}(A \mid A^{(t)}) \equiv \log(A) - \frac{k-r}{2}\log(V_1 + A) - \frac{E(S^{com}|\mathbf{y}^{obs}, A^{(t)})}{2(V_1 + A)} \tag{2.35}$$

E-step In this case the E-step is identical to that in the REML calculations of

Section 2.3.2.

M-step Denoting $S_* \equiv E(S^{com}|\mathbf{y}^{obs}, A^{(t)})$ found in the E-step. We have that

$$\frac{\partial Q_{adj}}{\partial A} = \frac{1}{A} - \frac{k-r}{2(A+V_1)} + \frac{S_*}{2(V_1+A)^2} \tag{2.36}$$

Solving $\frac{\partial Q_{adj}}{\partial A} = 0$ and letting $m = k-r-2$ gives two unique roots where the positive root equals

$$\hat{A}^{t+1} \equiv \frac{\sqrt{(2S_* - V_1(m-2))^2 + 8mV_1} - (2S_* - (m-2)V_1)}{2m}. \tag{2.37}$$

2.3.4 Data Augmentation Properties

The traditional EM procedure for the Normal-Normal hierarchical model is a well known result and has been studied at length elsewhere and so in this section we will focus on some of the properties of the data-augmented algorithms.

Theorem 2.3.1. *In the equal variance case, $V_i = V$ for all i such that $\mathbf{y}^{com} = \mathbf{y}^{obs} = \mathbf{y}$ the MLE, REML and ADMDA data-augmented algorithms all converge in one step.*

Proof. For the MLE procedure the E-step reduces to

$$\begin{aligned} &E(\mathbf{y}^{com} \mid \mathbf{y}^{obs}, \phi = \phi^{(0)}) = \mathbf{y} \\ &\sum_{i=1}^{k} E((y_i^{com} - x_i'\beta^{(1)})^2 = \sum_{i=1}^{k}(y_i - x_i'\beta^{(1)})^2 \end{aligned} \tag{2.38}$$

and thus the M-step reduces to the maximum likelihood result

$$\begin{aligned} \hat{\beta}^{(1)} &\equiv (X'X)^{-1}X'\mathbf{y} \\ \hat{A}^{(1)} &\equiv \max\left\{\frac{\sum_{i=1}^{k}(y_i - x_i'\beta^{(1)})^2}{k} - V_1, 0\right\}. \end{aligned} \tag{2.39}$$

For the REML procedure the E-step reduces to

$$S^* \equiv E(S^{com}|\mathbf{y}^{obs}, A^{(t)}) = \sum_{i=1}^{k}(y_i - x_i'\hat{\beta})^2 \tag{2.40}$$

□

where $\hat{\beta} \equiv (X'X)^{-1}X'\mathbf{y}$. Thus the M-step reduces to the REML solution

$$A^{(1)} \equiv \max\left\{\frac{S_*}{k-r} - V_1, 0\right\}. \tag{2.41}$$

For the ADM procedure the E-step is equivalent to the E-step for REML and thus the M-step reduces to the ADM maximum under equal variances

$$\hat{A}^{(1)} \equiv \frac{\sqrt{(2S_* - V_1(m-2))^2 + 8mV_1} - (2S_* - (m-2)V_1)}{2m}. \tag{2.42}$$

Theorem 2.3.2. *The data-augmented ADM algorithm is guaranteed to not decrease the adjusted observed log-likelihood after each iteration.*

Proof. Note that this proof is almost a direct consequence of the proof in Dempster

et al. (1977a). From Bayes rule we note that

$$p(y^{obs} \mid A) = \frac{p(y^{obs}, y^{mis} \mid A)}{p(y^{mis} \mid y^{obs}, A)} \tag{2.43}$$

and hence

$$A \times p(y^{obs} \mid A) = \frac{A \times p(y^{obs}, y^{mis} \mid A)}{p(y^{mis} \mid y^{obs}, A)} \tag{2.44}$$

where $A \times p(y^{obs} \mid A)$ is defined to be our adjusted observed data likelihood and $A \times p(y^{obs}, y^{mis} \mid A)$ is defined as our adjusted complete data likelihood. Therefore from (2.44) we see that

$$\log(A) + \log(p(y^{obs} \mid A)) = \log(A) + \log(p(y^{obs}, y^{mis} \mid A)) - \log(p(y^{mis} \mid y^{obs}, A)). \tag{2.45}$$

Taking expectations of both sides over the distribution of $y^{mis} \mid y^{obs}, A^{(t)}$ gives

$$\begin{aligned} \log(A) + \log(p(y^{obs} \mid A)) &= \log(A) + E(\log(p(y^{obs}, y^{mis} \mid A))) \\ &\quad - E(\log(p(y^{mis} \mid y^{obs}, A))). \\ &= Q(A \mid A^{(t)}) + H(A \mid A^{(t)}) \end{aligned} \tag{2.46}$$

where our adjusted Q-function is defined as

$$Q(A \mid A^{(t)}) \equiv \log(A) + E(\log(p(y^{obs}, y^{mis} \mid A))) \tag{2.47}$$

and

$$H(A \mid A^{(t)}) \equiv -E(\log(p(y^{mis} \mid y^{obs}, A))). \tag{2.48}$$

Note that from (2.46) that

$$\log(A^{(t)}) + \log(p(y^{obs} \mid A^{(t)})) = Q(A^{(t)} \mid A^{(t)}) + H(A^{(t)} \mid A^{(t)}) \tag{2.49}$$

is also true and hence

$$\begin{aligned}(\log(A) + \log(p(y^{obs} \mid A))) - (\log(A^{(t)}) + \log(p(y^{obs} \mid A^{(t)})))& \\ = Q(A \mid A^{(t)}) - Q(A^{(t)} \mid A^{(t)})& \\ + H(A \mid A^{(t)}) - H(A^{(t)} \mid A^{(t)}).&\end{aligned} \tag{2.50}$$

Note that our definition of our H function has not differed from the original proof in Dempster et al. (1977a) and hence via a consequence of Jensen's inequality

$$H(A|A^{(t)}) \geq H(A^{(t)} \mid A^{(t)}) \tag{2.51}$$

still holds true. Therefore

$$(\log(A)+\log(p(y^{obs} \mid A)))-(\log(A^{(t)})+\log(p(y^{obs} \mid A^{(t)}))) \geq Q(A \mid A^{(t)})-Q(A^{(t)} \mid A^{(t)}) \tag{2.52}$$

and hence maximizing our adjusted Q-function and finding an A to improve upon $Q(A \mid A^{(t)})$ over $Q(A^{(t)} \mid A^{(t)})$ will improve our adjusted log-likelihood by as least as much.

□

2.3.5 Performance Comparisons

In this section we present two simulation studies to investigate the performance properties of the various algorithms presented in Sections 2.3.2, 2.3.1 and 2.3.3. In each simulation the mean number of iterations (over 1000 simulations) to converge for each algorithm are calculated over a range of B_h values where

$$B_h \equiv \frac{V_h}{V_h + A} \tag{2.53}$$

and V_h is the harmonic mean of the variances V_i. It is noted that B_h has the nice property that it is also the harmonic mean of the shrinkage factors and was merely chosen for convenience to represent typical shrinkage. In order for a fair comparison to exist each algorithm was given the same initial starting values of $A^{(0)} = V_h$ and $\beta^{(0)} = 10$ regardless of the true values of A and β and the same convergence condition to stop when

$$Q(A^{(t+1)} \mid A^{(t)}) - Q(A^{(t)} \mid A^{(t)}) < 10^{-8} \tag{2.54}$$

or when the maximum number of iterations (set to 100) is reached.

The first simulation results are presented in Figures 2.1 and 2.2and it is for the scenario where $k = 10$, $\beta = 0$ and $V_i = 1$ for $i = 1 \ldots 5$ and $V_i = 2$ for $i = 6 \ldots 10$.

In Figure 2.2 the mean time is calculated by running the simulation on one core of a Intel(R) Core(TM) i7-2860QM CPU @ 2.50GHz. It is noted in Figures 2.1 and 2.2 that the ADMDA procedure performs well when compared to the other algorithms. The traditional REML procedure perform the worst on both metrics and this is largely

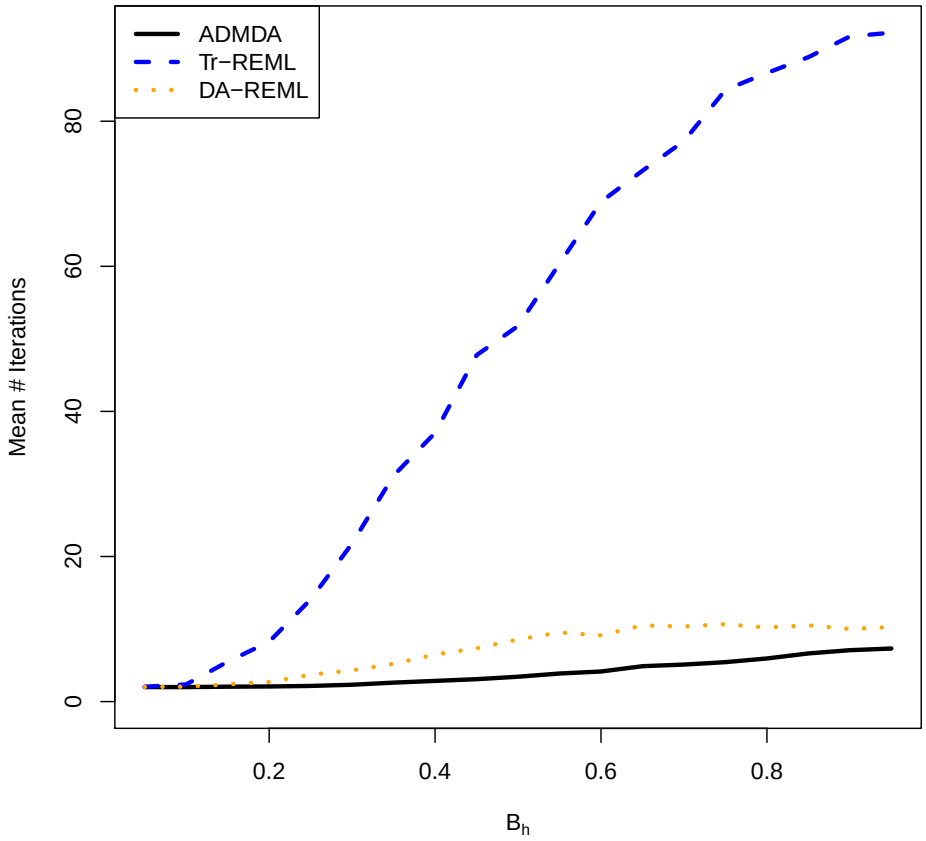

Figure 2.1: Mean Number of Iterations for the Two Group Variance Scenario

due to the fact that in the scenario when the REML procedure would give an estimate of $\hat{A} = 0$ the traditional EM algorithm creeps so slowly toward the mode that the maximum number of iterations is often hit. The data-augmented REML procedure, however, does not suffer this fate and it is has been observed in practice to produce $\hat{A} = 0$ for some datasets after only a few iterations. Either way an estimate of $\hat{A} = 0$ is undesirable. If the user of the algorithm does not realize this then it may be useful to receive a maximum iteration hit message from the traditional EM. However, for a more experienced user it's clear that knowing relatively early that an estimate $\hat{A} = 0$ has occurred would be more beneficial. Regardless, due to the reasons noted

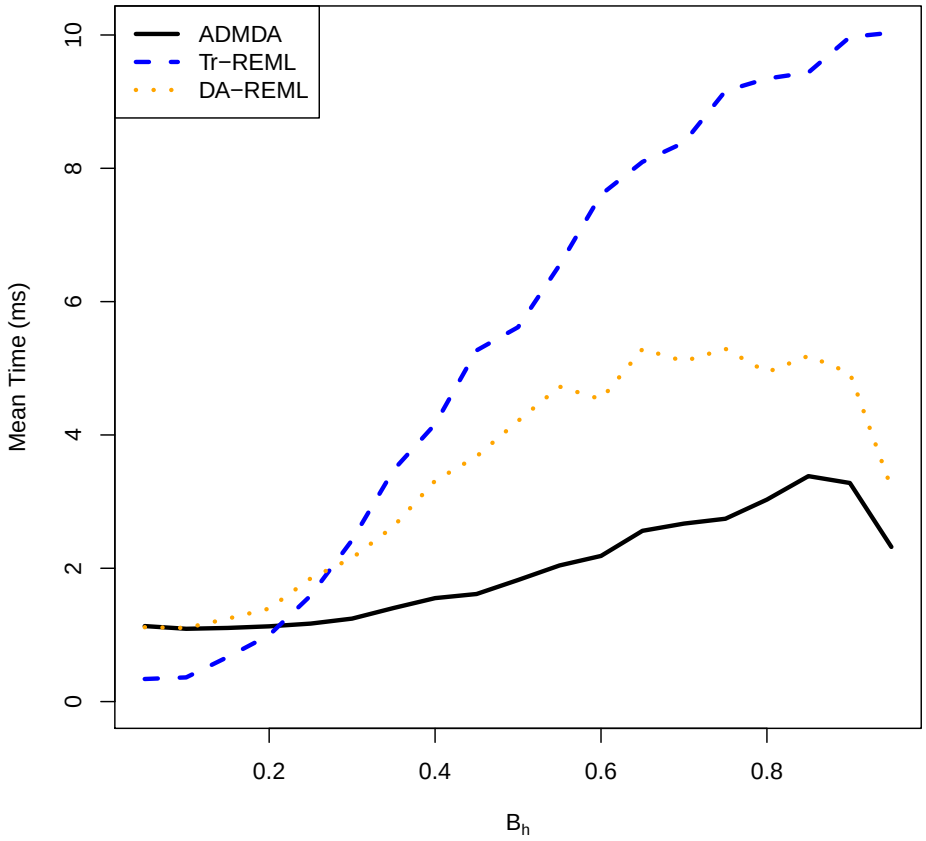

Figure 2.2: Times (ms) for the Two Group Variance Scenario

in Chapter 1 the ADMDA procedure is clearly the preferred algorithm.

The second simulation results are presented in Figure 2.3 and it is for the scenario where $k = 10$, $\beta = 0$ and $V_i = i$ for $i = 1 \dots k$. This simulation represents a scenario where the variances are very different and thus the data augmentation procedure should not perform as well due to the large fraction of missing data present in many of the groups.

Indeed by observing Figures 2.3 and 2.4 we see that the traditional REML procedure is worst in terms of mean number of iterations but beats the data-augmented REML algorithm on time. This is because in this example the DA-REML algorithm

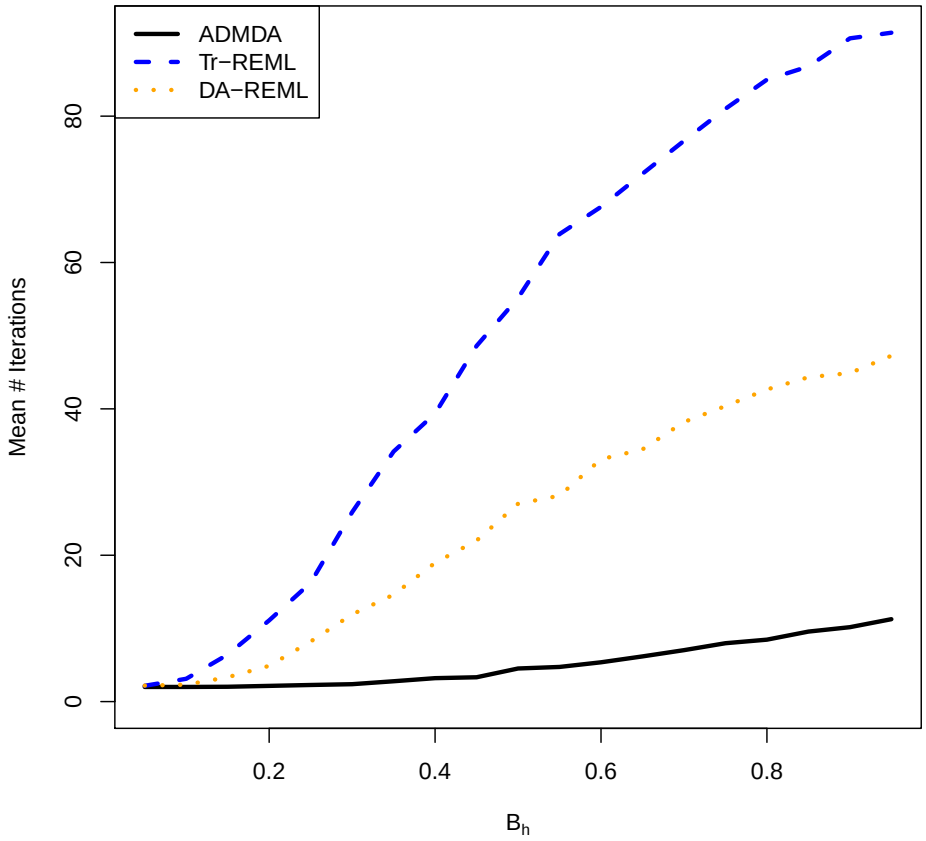

Figure 2.3: Mean Number of Iterations for the $V_i = i$ Scenario

is taking approximately twice as many iterations to converge than in the previous example due to the larger proportion of missing data. As the REML procedure involves more calculations at each iteration it thus takes more time (ms) to converge than the Tr-REML algorithm. The ADMDA still performs the best on both metrics, however, and this is due to the fact that the adjusted complete data likelihood gets pinned to the value of 0 when $A = 0$ regardless of the value of the positive valued random variable S_{com}. For most datasets this has the effect of producing a Q-function which is relatively quadratic and thus easy to maximize.

In Chapter 1 we examined how GRIMM performed for the hospital data (Section

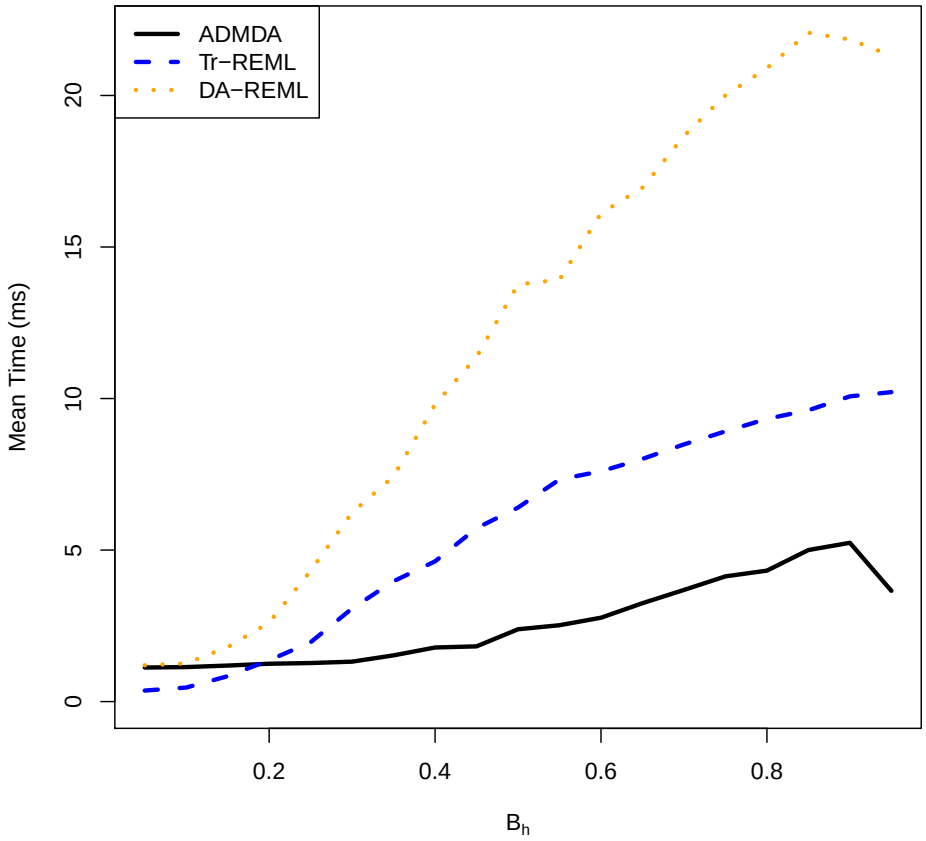

Figure 2.4: Times (ms) for the $V_i = i$ Scenario

1.7.2). In Figure 2.5 we revisit this dataset and demonstrate the converge of the ADMDA algorithm under the initial starting value of $A^{(0)} = V_h$.

In Figure 2.5 we see that convergence is very quick with the mode almost being reached after only a few iterations. Given that closed-form solutions exists for both the E-step and M-step for the ADMDA algorithm this leads to a computational efficient estimation procedure.

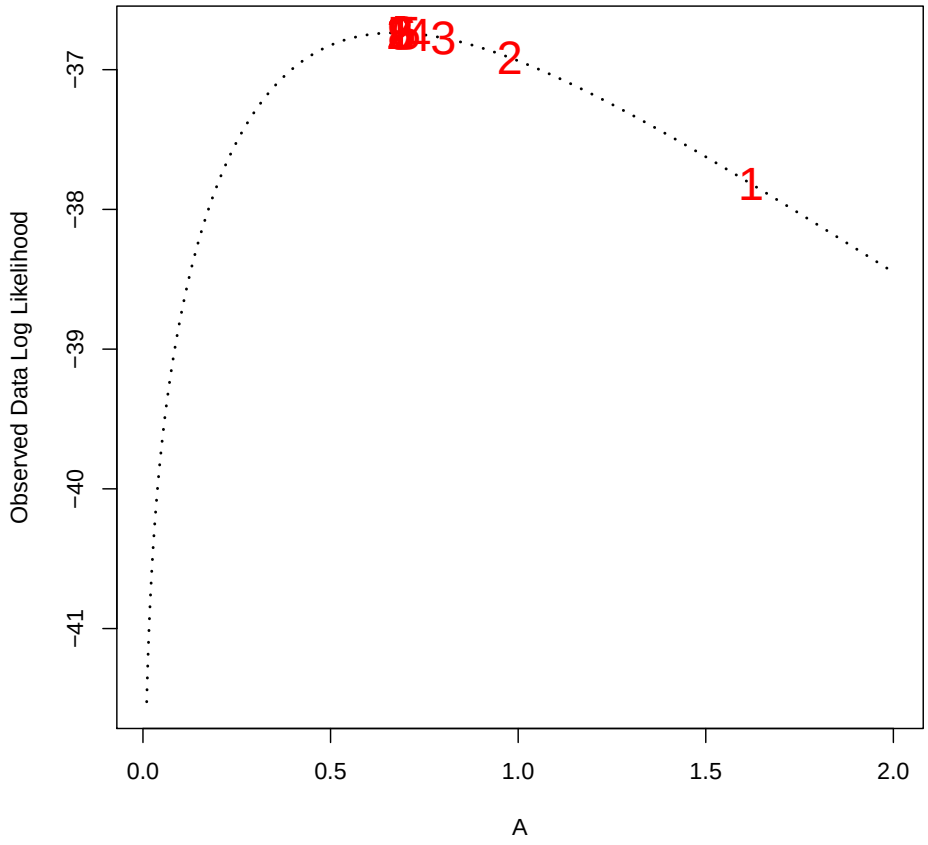

Figure 2.5: Convergence of ADMDA Algorithm for the Hospital Data

2.4 Gibbs Sampling Algorithms

Although point estimation is often all that is required sometimes it is necessary to make inferences about the entire posterior distribution of the random effects. Interestingly, the data augmentation scheme presented in Section 2.2 can be used in a Gibbs sampler to help improve the computational efficiency by reducing the autocorrelation between draws. In this section we present the traditional and data augmented Gibbs sampler and compare them on the schools and hospital example datasets.

2.4.1 Traditional Gibbs Sampler

Assuming flat priors on (β, A) as in Section 1.4 allows for the full conditional distributions to easily be obtained in a recognizable form allowing for an easy to implement Gibbs sampler. Firstly from (1.59) we see that

$$\theta_i \mid y_i, \beta, A \sim N((1 - B_i)y_i + B_i x_i^{'}\beta, V_i(1 - B_i)) \tag{2.55}$$

where $B_i \equiv V_i/(V_i + A))$. We can also easily derive

$$A \mid \beta, \theta \sim \text{Inv-Gam}(k/2 - 1, \sum_{i=1}^{k} \frac{(\theta_i - x_i^{'}\beta)^2}{2}) \tag{2.56}$$

and

$$\beta \mid \Theta, A \sim N_r(X\hat{\beta}, A(X^{'}X)^{-1}) \tag{2.57}$$

where $\hat{\beta} = (X'X)^{-1}X'\Theta$. This leads to Gibbs sampler described in Algorithm 2.1 with the number of draws equal to T and a burn-in of size N.

Algorithm 2.1 Traditional Gibbs Sampler

Initialize: $\beta^{(0)}$, $A^{(0)}$, T, N.
  for t in $1:T$ **do**
    for i in $1:k$ **do**
      $\theta_i^{(t)} \leftarrow$ Draw from $\theta_i \mid y_i, \beta^{(t-1)}, A^{(t-1)}$
    end for
    $\beta^{(t)} \leftarrow$ Draw from $\beta \mid \Theta^{(t)}, A^{(t-1)}$
    $A^{(t)} \leftarrow$ Draw from $A \mid \beta^{(t)}, \Theta^{(t)}$
  end for
Return: $A^{((N+1):T)}$, $\beta^{((N+1):T)}$, $\Theta^{((N+1):T)}$

2.4.2 Data Augmented Gibbs Sampler

Assuming flat priors on (β, A) as in Section 1.4 and utilizing the methodology present in Section 2.2 allows us to derive the full conditional distributions in a recognizable form.

$$\beta \mid \mathbf{y^{com}}, A \sim N(\hat{\beta}, (V_1 + A)(X'X)^{-1}) \tag{2.58}$$

$$y_i^{com} \mid y_i^{obs}, \beta, A \overset{ind}{\sim} N((1 - w_i B_i) y_i^{obs} + w_i B_i x_i' \beta, w_i V_1 + w_i^2 (V_i(1 - B_i))) \tag{2.59}$$

$$R = V_1 + A \mid \mathbf{y}^{com}, \beta \sim \text{Inv-Gamma}(\frac{k}{2} - 1, \frac{\sum_{i=1}^{k} (y_i^{com} - x_i'\beta)^2}{2}) \quad \text{where } R > V_1. \tag{2.60}$$

This leads to the following Gibbs sampler described in Algorithm 2.2 where the number of draws are equal to T and there is a burn-in of size N.

Algorithm 2.2 Data Augmented Gibbs Sampler

Initialize: $\mu_0^{(0)}$, $A^{(0)}$, T, N.
  for t in $1:T$ **do**
    for i in $1:k$ **do**
      $y_i^{com(t)} \leftarrow$ Draw from $y_i^{com} \mid y_i^{obs}, \mu_0^{(t-1)}, A^{(t-1)}$
    end for
    $\beta^{(t)} \leftarrow$ Draw from $\beta \mid (\mathbf{y^{com}})^{(t)}, A^{(t-1)}$
    $A^{(t)} \leftarrow 0$
    while $A^{(t)} \leq 0$ **do**
      $R \leftarrow$ Draw from $R \mid (\mathbf{y}^{com})^{(t)}, \beta^{(t)}$
      $A^{(t)} \leftarrow R - V_1$
    end while
  end for
Return: $A^{((N+1):T)}$, $\beta^{((N+1):T)}$

2.4.3 Performance Comparisons

In this section we compare the two Gibbs sampling algorithms presented in Section 2.4 for the school and hospital data first seen in Chapter 1.

In Figure 2.6 we note the results of running the Gibbs sampler for 100000 draws with a burn-in period of 50000 for the schools data. Note that for this example the data augmented Gibbs has less correlation between subsequent draws and thus is able to achieve a higher effective sample size (26178 versus 16682) after accounting for the correlation present between draws (Plummer et al. (2006)).

In Figure 2.7 we see similar results when running the two Gibbs samplers for 100000 draws with a burn-in period of 50000 on the hospital data. Note that for this

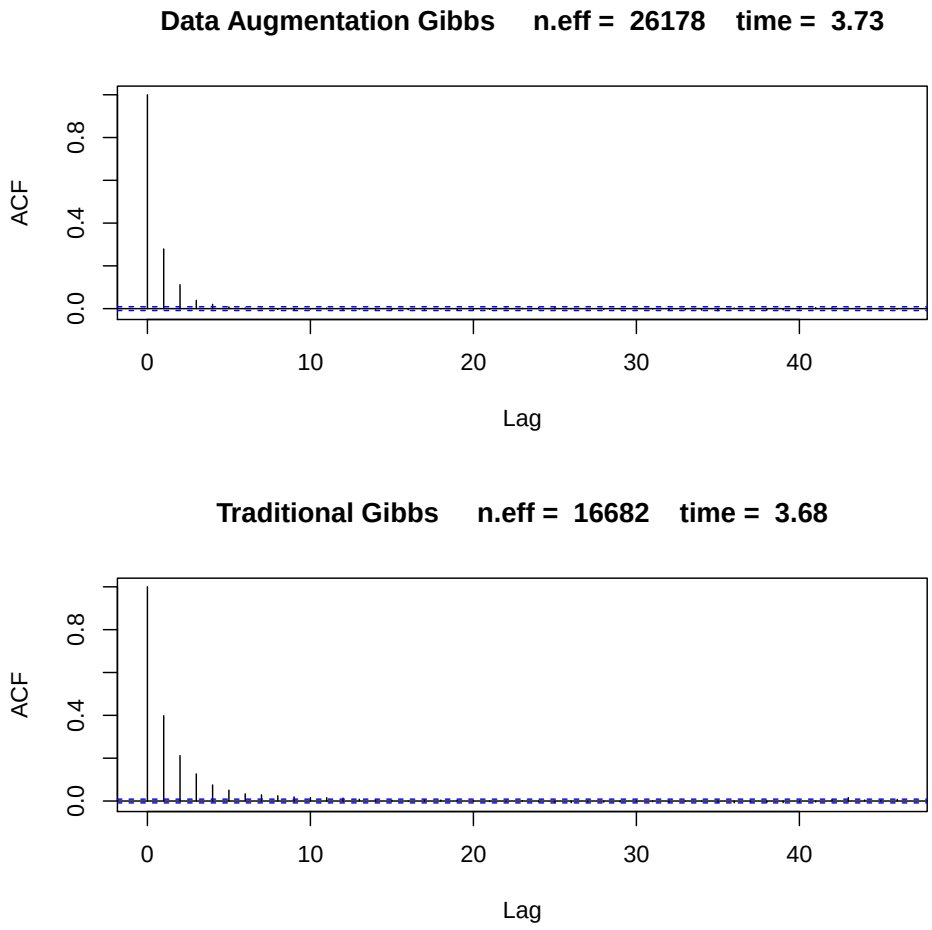

Figure 2.6: Autocorrelation Plot and Effective Sample Size for the Schools Data

example the data augmented Gibbs has much less autocorrelation and thus is able to produce over three times as many effective draws in less computation time.

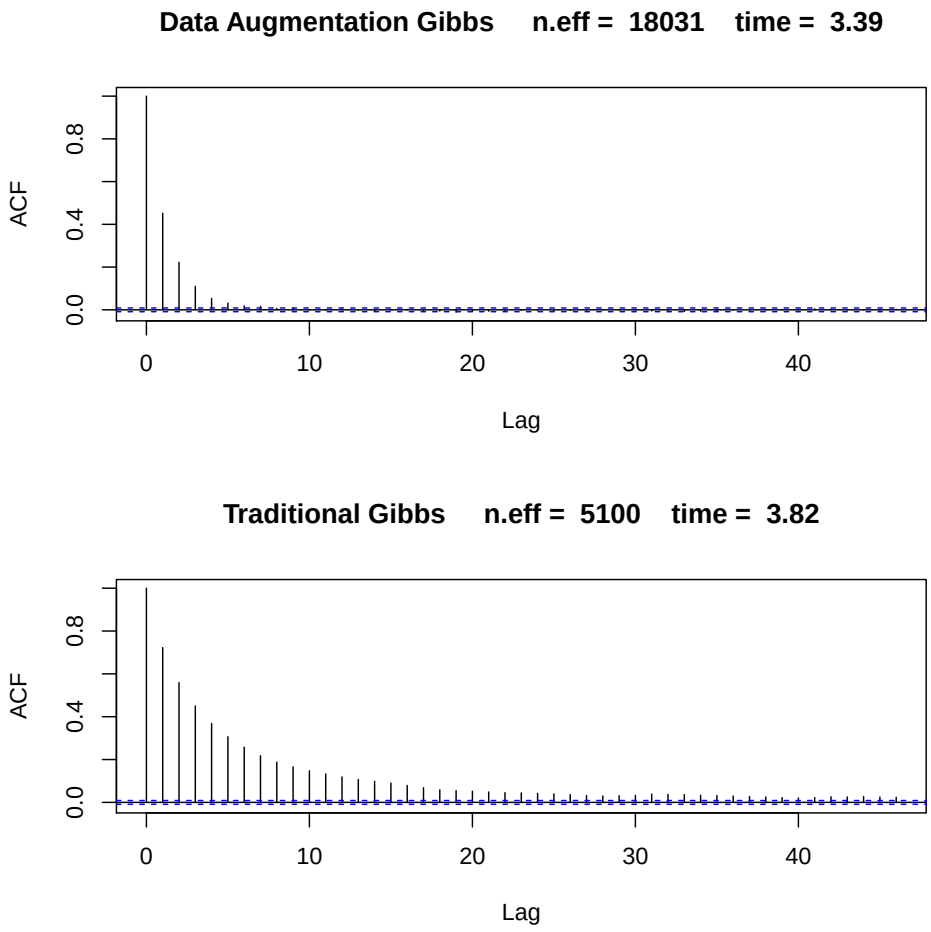

Figure 2.7: Autocorrelation Plot and Effective Sample Size for the Hospitals Data

2.5 Summary and Conclusions

In Section 2.2 we presented and derived some theoretical properties of a data augmentation scheme in order to fit the Normal-Normal hierarchical model. In 2.3 we examined some EM algorithms to fit the model via MLE and REML under two schemes of complete data which we dubbed traditional and data augmentation. In Section 2.3.3 we proposed the new ADMDA algorithm combining ideas from ADM and EM to produce a very reliable and fast algorithm as shown in Figures 2.1, 2.3 and 2.5.

In order to justify the use of ADMDA in Section 2.3.4 we derived some theoretical properties of the algorithm and in particular showed that for each iteration ADMDA will not decrease the adjusted log-likelihood.

The data augmentation scheme we examined was also shown to have benefits if undertaking MCMC. In Section 2.4 we presented a Gibbs algorithm based on this data augmentation scheme and compared it with an alternative Gibbs algorithm. For the examples presented we see that the Gibbs algorithm based on the data augmentation method provided one and a half to three times the effective number of draws.

The benefits of having a fast algorithm like ADMDA are numerous. In particular, it is noted in Figures 2.6 and 2.7 that it takes approximately 4 seconds to obtain enough draws to make reasonable inferences. By comparison, as seen in Figures 2.2 and 2.4, the ADMDA algorithm can fit the model with similar data at least 4000 times in the same time frame. The relative speed of ADMDA over MCMC allows for

method checking which requires the model be estimated numerous times, something which is impractical with MCMC.

In conclusion, the data augmentation method examined in this chapter can prove to be beneficial when conducting point estimation and when computing the posterior via MCMC. Particular note is made of the new ADMDA algorithm which augments the data and creates missing data. This data augmentation scheme coupled with ADM allows for a very quick computational method where the E-step and M-step are in closed form whilst providing the good frequency properties that assuming Stein's harmonic prior aims to supply.

Appendix A

Historical Perspective

In this chapter we present a historical perspective of the Normal-Normal hierarchical model. This chapter is included in the Appendix as it didn't seem appropriate to include in the main text but may provide the reader with a good background on some of the theories contained in the dissertation.

The earliest version of the Normal-Normal hierarchical model could be attributed to Charles Stein. Though it seems that Stein was not viewing the model in the context of a hierarchical model but proving the inadmissibility of the standard sample mean estimator for the mean of a multivariate Normal distribution. Stein (1956)

Stein considered the case of estimating $\Theta = (\theta_1, \ldots \theta_k)$ when

$$y_i \mid \theta_i \sim N(\theta_i, 1) \text{ for } i = 1 \ldots k. \tag{A.1}$$

Up until Stein's contributions, the usual estimator of θ_i was the sample mean $\hat{\theta}_i^{(sm)} = y_i$. Stein, however, shows that for $k \leq 2$ this estimator is admissible but for $k \geq 3$ it is inadmissible with respect to the loss function of the sum of squares of the errors

$$L(\Theta, \hat{\Theta}) = \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2. \tag{A.2}$$

To prove this, Stein considered a spherically symmetric estimator of the form

$$\hat{\theta}_i^{(S)} = \left(1 - \frac{b}{a + ||\mathbf{y}||^2}\right) y_i \tag{A.3}$$

where $a, b > 0$ and $||\mathbf{y}||^2 = \mathbf{y}'\mathbf{y}$ and proved that this alternative estimator had smaller risk under a squared error loss function when $a \to \infty$ and $0 < b < 2(k-2)$

Stein extended his 1956 result and partnered up with Willard James in 1961 to produce what is now known as the James-Stein estimator. James and Stein (1961) considered a special case of the spherically symmetric estimator in (A.3)

$$\hat{\theta}_i^{(JS)} = \left(1 - \frac{(k-2)}{||\mathbf{y}||^2}\right) y_i \tag{A.4}$$

that dominates the sample mean estimator in terms of expected squared error loss, $R(\Theta, \hat{\Theta}) = E(|\Theta - \hat{\Theta}|^2)$, otherwise known as risk.

Proof. We would like to prove $R(\Theta, \hat{\Theta}^{(JS)}) < R(\Theta, \hat{\Theta}^{(sm)})$ where $\hat{\Theta}^{(sm)}$ is the sample

mean estimator. Firstly, note that

$$R(\Theta, \hat{\Theta}^{(sm)}) = E||\Theta - \mathbf{y}||^2 = \sum_{i=1}^{k} E(\theta_i - y_i)^2 = k \tag{A.5}$$

and as in (A.3) consider an estimator of the form

$$\hat{\theta_i}^{(S)} = \left(1 - \frac{b}{a + ||\mathbf{y}||^2}\right) y_i \tag{A.6}$$

where $a, b > 0$. Then

$$\begin{aligned} R(\Theta, \hat{\Theta}^{(S)}) &= E\left(||\Theta - (1 - \frac{b}{a + ||\mathbf{y}||^2})\mathbf{y}||^2\right) & \text{(A.7)} \\ &= E(||\Theta - \mathbf{y}||^2) + E\left(\frac{2b(\Theta - \mathbf{y})'\mathbf{y}}{a + ||\mathbf{y}||^2}\right) + E\left(\frac{b^2||\mathbf{y}||^2}{(a + ||\mathbf{y}||^2)^2}\right) & \text{(A.8)} \end{aligned}$$

Unlike Stein's original proofs from Stein (1956) and James and Stein (1961) here we utilize a lemma from his papers on the unbiased estimation of risk Stein (1973) and Stein (1981). The lemma states that

$$E(g(X)(X - \mu)) = \sigma^2 E(g'(X)) \tag{A.9}$$

where $X \sim [\mu, \sigma^2]$. Let

$$g_i(\mathbf{y}) = \frac{y_i}{a + ||\mathbf{y}||^2} \tag{A.10}$$

then

$$g_i'(\mathbf{y}) = \frac{1}{a + ||\mathbf{y}||^2} - \frac{2y_i^2}{(a + ||\mathbf{y}||^2)^2}. \tag{A.11}$$

Hence

$$\frac{2b(\Theta - \mathbf{y})'\mathbf{y}}{a + ||\mathbf{y}||^2} = \frac{\sum_{i=1}^{k}(\theta_i - y_i)y_i}{a + ||\mathbf{y}||^2} \tag{A.12}$$

$$= -\frac{k}{a + ||\mathbf{y}||^2} + \frac{2||\mathbf{y}||^2}{(a + ||\mathbf{y}||^2)^2} \tag{A.13}$$

Then

$$R(\Theta, \hat{\Theta}^{(S)}) = E(||\Theta - \mathbf{y}||^2) - 2bE\left(\frac{k}{a + ||\mathbf{y}||^2} - \frac{2||\mathbf{y}||^2}{(a + ||\mathbf{y}||^2)^2}\right) + E\left(\frac{b^2||\mathbf{y}||^2}{(a + ||\mathbf{y}||^2)^2}\right) \tag{A.14}$$

and letting $a \to 0$ we see that

$$\lim_{a \to 0} R(\Theta, \hat{\Theta}^{(S)}) = k - 2b(k-2)E\left(\frac{1}{||\mathbf{y}||^2}\right) + b^2 E\left(\frac{1}{||\mathbf{y}||^2}\right) \tag{A.15}$$

and that this quadratic is minimized when $b = (k-2)$ hence giving the optimal estimator $\hat{\theta}_i^{(JS)} = \left(1 - \frac{(k-2)}{||\mathbf{y}||^2}\right) y_i$ with risk

$$R(\Theta, \hat{\Theta}^{(JS)}) = k - (k-2)^2 E\left(\frac{1}{||\mathbf{y}||^2}\right) < k. \tag{A.16}$$

□

It is noted that for $(k-2) < ||\mathbf{y}||^2$ then this estimator can be thought of as shrinking the sample mean, y_i, toward 0. When $(k-2) > ||\mathbf{y}||^2$ then this estimator overshoots the mark and $\hat{\theta}_i^{(JS)}$ has an opposite sign than y_i. To rectify this Stein

suggests

$$\hat{\theta}_i^{(JS)} = \left(1 - \min\left\{1, \frac{(k-2)}{||\mathbf{y}||^2}\right\}\right) y_i. \tag{A.17}$$

Bradley Efron and Carl Morris extended Stein's 1961 result in their 1973 paper by taking an empirical Bayes approach. As in James and Stein (1961) Efron and Morris assume (A.1) but make a point to note that if $y_i \sim N(\theta_i, \sigma^2/n)$ where σ is known then a change of scale transforms σ^2/n to the more convenient value of 1 Efron and Morris (1973).

Efron and Morris are the first to formally recognize the James-Stein estimator as an estimator in the framework of a hierarchical model. They assume

$$y_i \overset{ind}{\sim} N(\theta_i, 1) \tag{A.18}$$

$$\theta_i \overset{i.i.d}{\sim} N(0, A) \text{ for } i = 1 \ldots k \tag{A.19}$$

and note that

$$\theta_i \mid \mathbf{y}, A \overset{ind}{\sim} N((1-B)y_i, (1-B)) \tag{A.20}$$

where

$$B = \frac{1}{A+1} \tag{A.21}$$

is known as a shrinkage factor. If B is assumed to be known this leads to the Bayes estimator

$$\hat{\theta}_i = (1-B)y_i. \tag{A.22}$$

Efron and Morris draw the connection that

> one can take the assumption that $\theta_i \sim N(0, A)$ seriously, but stop short of full Bayesianhood by assuming that A is unknown and must be estimated from the data. That is, one can be an empirical Bayesian.

The James-Stein estimator can be seen to approximate the Bayes rule when B is unknown and is estimated as

$$\hat{B}_{js} = \frac{(k-2)}{||\mathbf{y}||^2}. \tag{A.23}$$

As in James and Stein (1961) they note that by construction $0 < B \leq 1$, however, if $(k-2) > ||\mathbf{y}||^2$ then $\hat{B}_{js} > 1$. To rectify this they suggest an estimator,

$$\hat{B}_{em} = \min\{\hat{B}_{js}, 1\}, \tag{A.24}$$

that uniformly improves upon $\hat{B}_{js}$.

In this paper Efron and Morris also make note that in the unequal variance case that transforming and estimating under the canonical case $\tilde{y}_i \sim N(\theta_i, 1)$ leads to each y_i being shrunk by the same factor, even though each group has a different amount of information as indicated by the differing variances. They suggest methods of estimation under the original scale and continue their ideas in their later papers.

In Efron and Morris (1975) Efron and Morris review Stein's estimator and the unequal variance generalization and apply them to predict baseball averages, estimate the prevalence of toxoplasmosis in El Salvador and estimate the exact size of Pearson's chi-square test with results from a computer simulation. In addition to the applications listed in Efron and Morris (1975) a review of a current methods, aptly titled *Stein's Paradox in Statistics*, was published by Efron and Morris (1977). The

review provides a summary of methods used at the time and is a great starting point for those wishing to learn more about Stein and his contributions to statistics.

In Morris (1983a) Morris reviews the state of multi-parameter shrinkage estimators in both the equal and unequal variance setting. Morris extends the simple model to include covariates and estimates Ty Cobb's true batting average for the period 1905-1928. In the unequal variance case Morris suggests an iterative technique, based on Bayes' theory, for estimating A. This estimator is shown to have held up in various computer simulations and is related to estimators in Fay and Herriot (1979), Carter and Rolph (1974) and Dempster et al. (1977b). Morris notes that in the case of when the prior mean is known then the MLE could also be used but when the prior mean is unknown it is better to use restricted maximum likelihood estimation (REML).

Shortly after the work of Stein and during the contributions of Efron and Morris there is an explosion in the use of Stein's result in practical applications. Carter and Rolph (1974) extend the James-Stein result to the unequal variance case and apply the model to estimate the probability that a fire alarm reported from a particular street box signals a fire rather than a false alarm or other emergency. They recommend an estimation method based on the method of moments and propose an algorithm to estimate the second level variance parameter A. Fay and Herriot (1979) then extend the results of Carter and Rolph (1974) to estimate the income of small places (population less than 1000) and consider the case of when covariates are present for the equal and unequal variance cases.

Although they did not assume a hierarchical model it is important to note the

contributions of both Hudson (1974) and Berger (1976), however, independently considered a more general unequal variance problem of finding an estimator for $\Theta = (\theta_1, \ldots, \theta_k)$ under quadratic loss $L(\Theta, \delta) = (\delta - \Theta)^t Q(\delta - \Theta)$ where Q is a $k \times k$ weighting matrix that is known and positive definite.

Even into the modern day the estimation of the Normal-Normal hierarchical model is still an area of contention and new methods are often proposed. There is still much debate about optimal estimators and their properties and as recent as 2012 a possible solution was put forward in a recent paper by Xie et al. (2012). In this paper a new estimator is proposed for the unequal variance case based on Stein's unbiased estimate of risk (SURE) under the sum of squared error loss.

Whilst not able to provide a full summary of all the methods we highlight the results of a few important contributions. Morris and Tang (2011) utilize a procedure from Morris (1988) that suggests improvements to approximating univariate probability densities based on only calculating two derivatives. Morris and Tang (2011) name this procedure as Adjustment for Density Maximization and utilize it to fit the Normal-Normal hierarchical model. A closer look at this procedure is examined in Section 1.6.1.

To further learn about the commonly used shrinkage estimator the reader is pointed to the excellent review of shrinkage estimation in multilevel Normal models in Morris and Lysy (2012).

Appendix B

Hospital Unequal Variances Example

In this chapter we present the results of comparing GRIMM, MLE, REML and MCMC estimation methods for the unequal variances hospital example.

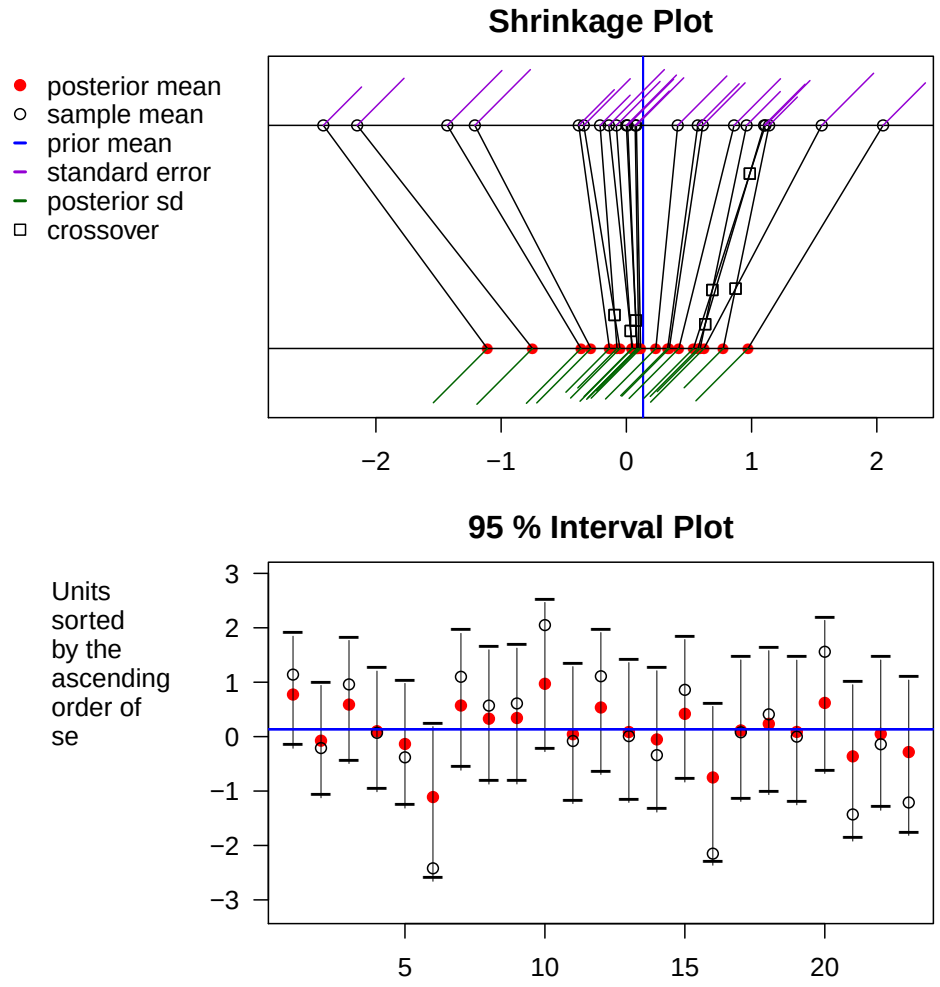

Figure B.1: Plot of the results of the analysis of the equal variances hospital data with GRIMM

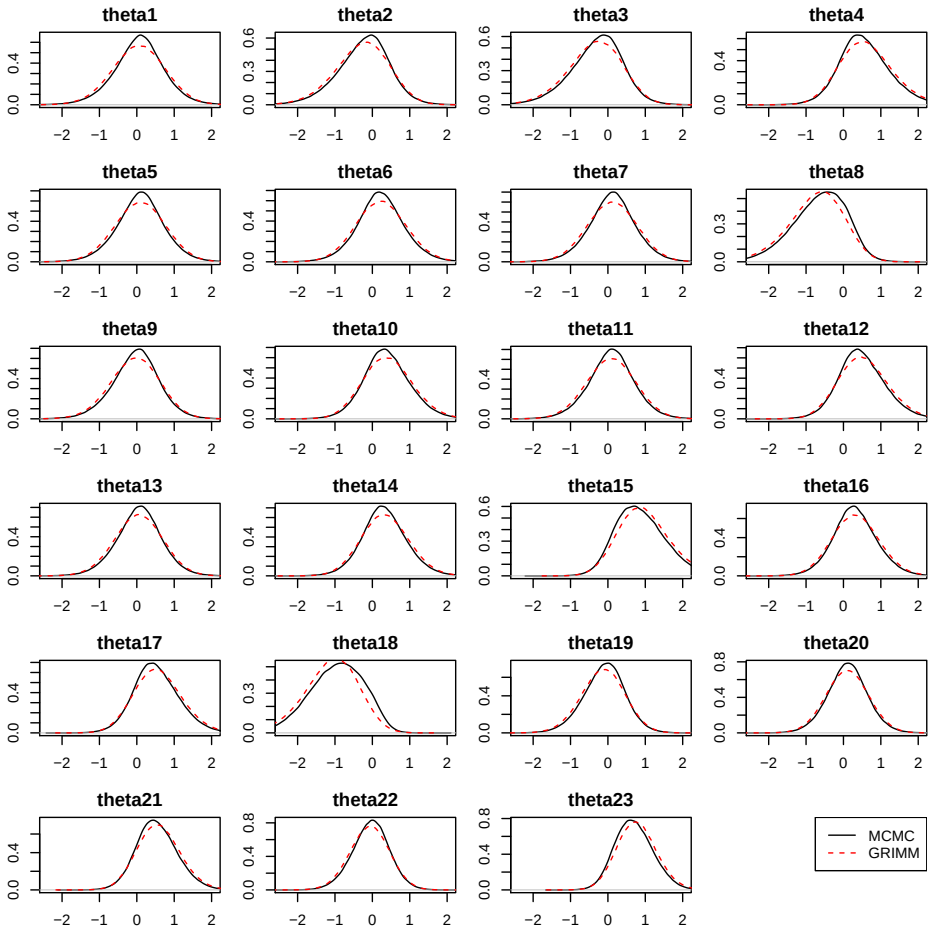

Figure B.2: Plot of the comparison between GRIMM and MCMC for the unequal variances hospital data

i	y_i	$\sqrt{V_i}$	$\hat{B}_i^{(M)}$	$\hat{B}_i^{(G)}$	$\hat{\theta}_i^{(M)}$	$\hat{\theta}_i^{(G)}$	$s_i^{(M)}$	$s_i^{(G)}$	$\hat{\gamma}_i^{(M)}$	$\hat{\gamma}_i^{(G)}$
1	-0.14	1.22	0.72	0.69	0.06	0.05	0.67	0.70	-0.10	-0.06
2	-1.21	1.22	0.72	0.69	-0.24	-0.28	0.70	0.73	-0.37	-0.30
3	-1.43	1.20	0.71	0.68	-0.31	-0.36	0.71	0.73	-0.41	-0.33
4	1.56	1.14	0.69	0.66	0.57	0.62	0.68	0.72	0.36	0.31
5	0.00	1.10	0.68	0.64	0.10	0.09	0.64	0.68	-0.06	-0.04
6	0.41	1.08	0.67	0.63	0.23	0.24	0.64	0.67	0.07	0.07
7	0.08	1.04	0.66	0.62	0.12	0.11	0.63	0.66	-0.03	-0.02
8	-2.15	1.03	0.66	0.61	-0.65	-0.75	0.73	0.74	-0.48	-0.43
9	-0.34	1.02	0.65	0.61	-0.03	-0.05	0.63	0.66	-0.16	-0.12
10	0.86	1.02	0.65	0.61	0.39	0.42	0.63	0.67	0.21	0.18
11	0.01	1.01	0.65	0.60	0.10	0.08	0.62	0.66	-0.06	-0.03
12	1.11	0.98	0.64	0.59	0.50	0.54	0.63	0.66	0.28	0.23
13	-0.08	0.96	0.63	0.58	0.06	0.04	0.61	0.64	-0.09	-0.06
14	0.61	0.93	0.61	0.56	0.32	0.34	0.60	0.64	0.14	0.11
15	2.05	0.93	0.61	0.56	0.88	0.97	0.67	0.70	0.42	0.36
16	0.57	0.91	0.60	0.55	0.31	0.33	0.59	0.63	0.13	0.10
17	1.10	0.90	0.60	0.55	0.52	0.57	0.61	0.64	0.27	0.22
18	-2.42	0.84	0.57	0.51	-0.96	-1.11	0.73	0.72	-0.35	-0.36
19	-0.38	0.78	0.54	0.48	-0.10	-0.14	0.56	0.58	-0.18	-0.11
20	0.07	0.75	0.52	0.46	0.11	0.10	0.54	0.57	-0.03	-0.01
21	0.96	0.74	0.52	0.45	0.54	0.59	0.55	0.58	0.24	0.18
22	-0.21	0.66	0.47	0.39	-0.04	-0.07	0.50	0.53	-0.13	-0.06
23	1.14	0.62	0.44	0.36	0.70	0.77	0.51	0.53	0.24	0.16

Table B.1: Results of the analysis of the unequal variances hospital data with MCMC(M) and GRIMM(G)

Appendix C

Data Augmentation Based Exact Estimator

In this chapter we tentatively propose a method to obtain a posterior mean estimate of the shrinkage factor, B_1, based on the data augmentation scheme in Chapter 2. This proposal, however, is an area of active research and as such we did not feel comfortable including it in the main text until we are confident of its properties and viability.

In (1.41) we derived an estimator for the equal variance case. We can therefore define an estimator based on the complete data, $\hat{B}_1^{com}$

$$\hat{B}_1^{com} \equiv E(B_1 \mid S^{com}) = \frac{(k-r-2)V_1}{S^{com}} \times \frac{P(\chi^2_{k-r} < S^{com}/V_1)}{P(\chi^2_{k-r-2} < S^{com}/V_1)}. \qquad \text{(C.1)}$$

Note this estimator is a function of our complete data and is a random variable. An appropriate estimator of $\hat{B}_1^{com}$ based only on observed data would then be the solution to

$$\hat{B}_1^{obs} \equiv E(\hat{B}_1^{com} \mid \mathbf{y}^{obs}). \tag{C.2}$$

We see can that $\hat{B}_1^{com}$ is not linear in S^{com} and as such taking the expectation is no easy feat. It is suggested that this may be solved using the multiple imputation Algorithm C.1.

Algorithm C.1 Data Augmented Exact Estimator

Initialize: $A^{(0)}$, T, N.
  for t in $1:T$ **do**
    for i in $1:k$ **do**
      $y_i^{com(t)} \leftarrow$ Draw from $y_i^{com} \mid y_i^{obs}, A^{(t-1)}$
    end for
    Estimate $\hat{B}_1^{com}$ from complete data.
    $A^{(t)} \leftarrow \frac{V_1(1-\hat{B}_k^{com})}{\hat{B}_1^{com}}$
  end for
Return: $A^{((N+1):T)}$

Note that this algorithm will not produce draws from the full joint posterior for A but rather after a sufficient burn in period will produces draws of A about the mean of $\hat{B}_1^{com}$. The convergence and other properties of this type of algorithm have not been investigated and thus this only a suggestion and is of need of further research.

Bibliography

Alderman, D. L. and Powers, D. E. The effects of special preparation on sat-verbal scores. *American Educational Research Journal*, 17(2):pp. 239–251, 1980. ISSN 00028312. URL http://www.jstor.org/stable/1162485.

Azzalini, A. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005. ISSN 1467-9469. doi: 10.1111/j.1467-9469.2005.00426.x. URL http://dx.doi.org/10.1111/j.1467-9469.2005.00426.x.

Berger, J. Admissible minimax estimation of a multivariate Normal mean with arbitrary quadratic loss. *The Annals of Statistics*, 4(1):223–226, 1976.

Berger, J. O., Strawderman, W., and Tang, D. Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Annals of statistics*, pages 606–646, 2005.

Brillinger, D. The calculation of cumulants via conditioning. *Annals of the Institute of Statistical Mathematics*, 21(1):215–218, 1969. ISSN 0020-3157. doi: 10.1007/BF02532246. URL http://dx.doi.org/10.1007/BF02532246.

Carter, G. and Rolph, J. Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, 69(348):pp. 880–885, 1974. ISSN 01621459. URL http://www.jstor.org/stable/2286157.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1):1–38, 1977a.

Dempster, A. P., Schatzoff, M., and Wermuth, N. A simulation study of alternatives to ordinary least squares: Rejoinder. *Journal of the American Statistical Association*, 72(357):pp. 104–106, 1977b. ISSN 01621459. URL http://www.jstor.org/stable/2286916.

Efron, B. and Morris, C. Stein's estimation rule and its competitors–An empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):pp. 117–130, 1973. ISSN 01621459. URL `http://www.jstor.org/stable/2284155`.

Efron, B. and Morris, C. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Efron, B. and Morris, C. Stein's paradox in statistics. *Scientific American*, 236(5): 119–127, 1977.

Fay, R. and Herriot, R. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):pp. 269–277, 1979. ISSN 01621459. URL `http://www.jstor.org/stable/2286322`.

Hudson, H. M. *Empirical Bayes estimation*. Department of Statistics, Stanford University., 1974.

James, W. and Stein, C. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

Kelly, J., Tak, H., and Morris, C. *Rgbp: Gaussian, Poisson, and Binomial Hierarchical Modeling*, 2014a. URL `http://cran.r-project.org/web/packages/Rgbp/index.html`. R package version 1.0.6.

Kelly, J., Tak, H., and Morris, C. Rgbp: An R package for Gaussian, Poisson, and Binomial hierarchical modeling. Technical report, Harvard University, 2014b.

Michalak, S. and Morris, C. Posterior Propriety for Hierarchical Models with Log-Likelihoods that Have Norm Bounds. Technical report, Los Alamos National Laboratory and Harvard University, 2014.

Morris, C. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983a. URL `http://www.jstor.org/stable/2287098`.

Morris, C. Parametric Bayes confidence intervals. In *Scientific Inference, Data Analysis, and Robustness*, volume 48. Academic Press, 1983b.

Morris, C. Approximating posterior distributions and posterior moments. *Bayesian statistics*, 3:327–344, 1988.

Morris, C. and Lock, K. Unifying the named natural exponential families and their relatives. *The American Statistician*, 63(3):247–253, 2009.

Morris, C. and Lysy, M. Shrinkage estimation in multilevel Normal models. *Statistical Science*, 27(1):115–134, 2012.

Morris, C. and Tang, R. Estimating random effects via adjustment for density maximization. *Statistical Science*, 26(2):271–287, 2011.

Plummer, M., Best, N., Cowles, K., and Vines, K. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006. URL `http://CRAN.R-project.org/doc/Rnews/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Rubin, D. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):pp. 377–401, 1981. ISSN 03629791. URL `http://www.jstor.org/stable/1164617`.

SAS Institute Inc. *SAS/STAT Software, Version 9.3*. Cary, NC, 2011. URL `http://www.sas.com/`.

Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

Stein, C. M. Estimation of the mean of a multivariate Normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics*, pages 345–381, 1973.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

Tang, R. *Fitting and evaluating certain two-level hierarchical models*. PhD thesis, Harvard University, 2002.

Van Dyk, D. and Tang, R. The one-step-late pxem algorithm. *Statistics and Computing*, 13(2):137–152, 2003. ISSN 0960-3174. doi: 10.1023/A:1023256509116.

Xie, X., Kou, S., and Brown, L. SURE Estimates for a Heteroscedastic Hierarchical Model. *Journal of the American Statistical Association*, 107(500), 2012.